Outliers may not be automatically removed.

Karch, J.D.

AMERICAN PSYCHOLOGICAL ASSOCIATION

Journal of Experimental Psychology: General

ISSN: 0096-3445

2023, Vol. 152, No. 6, 1735–1753
https://doi.org/10.1037/xge0001357

Outliers May Not Be Automatically Removed

Julian D. Karch
Methodology and Statistics Department, Institute of Psychology, Leiden University

Researchers often remove outliers when comparing groups. It is well documented that the common practice of removing outliers within groups leads to inflated Type I error rates. However, it was recently argued by André (2022) that if outliers are instead removed across groups, Type I error rates are not inflated. The same study discusses that removing outliers across groups is a specific case of the more general concept of hypothesis-blind removal of outliers, which is consequently recommended. In this paper, I demonstrate that, contrary to this advice, hypothesis-blind outlier removal is problematic. Specifically, it almost always invalidates confidence intervals and biases estimates if there are group differences. It moreover inflates Type I error rates in certain situations, for example, when variances are unequal and data nonnormal. Consequently, a data point may not be removed solely because it is deemed an outlier, whether the procedure used is hypothesis-blind or hypothesis-aware. I conclude by recommending valid alternatives.

Public Significance Statement
Removing outliers is widespread in psychological research. This study demonstrates that removing a data point solely because it is an outlier is problematic. As a remedy, robust techniques are recommended.

Keywords: outlier removal, outlier exclusion, outlier detection, outlier analysis, outlier rejection

Supplemental materials: https://doi.org/10.1037/xge0001357.supp

Psychological data sets typically contain erroneous data that must be dealt with appropriately. Erroneous data can be caused by several factors, from equipment failure and misunderstanding of instructions to errors during data input. It is crucial to remove or correct erroneous data for many reasons, including the sensitivity of standard techniques, such as the *t*-test or analysis of variance (ANOVA), to erroneous data (Wilcox, 2016). Sometimes there is a well-defined procedure to detect erroneous data. For example, for reaction time data, responses faster than a certain threshold are considered impossibly fast (André, 2022) and are therefore removed. Often, however, no such well-defined procedure is available.

A remedy used by many researchers is to focus on identifying outliers instead. An *outlier* is a data point that is "too far removed" from the rest of the data (Pincus and Lewis, 1994). The appeal of focusing on outliers is that no external knowledge is needed. The observed data suffice. There are many outlier identification techniques, which essentially differ only in how they measure the distance of a data point from the rest of the data. The *Z*-score approach is the most popular in psychology (Bakker and Wicherts, 2014). First, the sample mean $\bar{X}$ and sample standard deviation *SD* are calculated. The absolute *z*-score of a data point X_i is then $Z_i = |X_i - \bar{X}|/SD$. A cutoff beyond which data points are considered outliers then needs to be chosen. Common values are 2 and 3 (Bakker and Wicherts, 2014).

While statistical textbooks aimed at psychological researchers agree that identifying outliers is essential, they have differing views on what should be done next. For instance, Howitt and Cramer (2007) recommended that "it is important to eliminate them [outliers] because they can be so misleading" (p. 28). In contrast, Stevens (2012) warned against removing a data point merely because it is an outlier. It is therefore not surprising to find considerable heterogeneity in how psychological researchers handle outliers (André, 2022; Bakker and Wicherts, 2014). Importantly, removing outliers to safeguard standard techniques is still widespread (André, 2022; Bakker and Wicherts, 2014).

However, many publications within the field of mathematical statistics warn against the practice of removing outliers to safeguard standard techniques, as it leads to underestimation of standard errors and, in turn, inflated Type I error rates (Wilcox, 1998). Bakker and Wicherts (2014) investigated the severity of this effect in a simulation study tailored to mirror psychological data. Their results show that automatically removing outliers with the intent to safeguard the two-sample *t*-test leads to Type I error rates as high as 20%, and they consequently warn against it.

In a recent article in the *Journal of Experimental Psychology: General*, André (2022) extended this work by distinguishing between identifying outliers within and across groups. Identifying outliers within groups involves computing the distance of each point from the rest of the data separately for each group. Conversely, when outliers are identified across groups, the distances are computed by taking into account all the data. For example, identifying outliers within groups involves computing *z*-scores based on

This article was published Online First April 27, 2023.
Julian D. Karch https://orcid.org/0000-0002-1625-2822
Julian D. Karch confirms sole responsibility for the following roles: study conceptualization and design, analysis and interpretation of results, and manuscript preparation.
Correspondence concerning this article should be addressed to Julian D. Karch, Methodology and Statistics Department, Institute of Psychology, Leiden University, PO Box 9555, Leiden, 2300 RB, The Netherlands. Email: j.d.karch@fsw.leidenuniv.nl

group-specific means and standard deviations. In contrast, the overall mean and standard deviation are used when identifying outliers across groups. André argued that inflated Type I error rates, as found by Bakker and Wicherts (2014), are limited to within-group outlier removal. They backed up this assertion with simulation studies, where Type I error rates were never inflated when outliers were removed across groups. André additionally discussed that removing outliers across groups is a particular case of a more general concept: If the outlier removal procedure is blind to the tested hypothesis, it does not inflate Type I error rates. They conclude by recommending hypothesis-blind outlier removal as a straightforward approach to safeguard standard techniques.

In this paper, I demonstrate that, contrary to this advice, hypothesis-blind outlier removal is not a valid approach to safeguard standard techniques. This practice generally leads to biased parameter estimates and invalid confidence intervals when the null hypothesis is false. Problems are less severe when the null hypothesis is true, but inflated Type I error rates, and biased parameter estimates still occur in certain situations. I showcase these problematic properties of hypothesis-blind outlier removal by means of simulation studies for the same examples used by André (2022) to argue in favor of hypothesis-blind outlier removal: (a) comparing two groups and (b) regression modeling. I conclude by recommending valid alternatives, which follow the general principle of modern, robust statistics: instead of removing outliers, use statistical techniques that are robust against outliers.

Null Hypothesis False: Removing Outliers Across Groups Invalidates Difference Estimates and Confidence Intervals

This section demonstrates that removing outliers across groups almost certainly biases parameter estimates and invalidates confidence intervals if the values in one group indeed tend to be larger.

Design

To highlight the pervasiveness of the problem, I used a design very similar to the one used by André (2022). I will first describe all the aspects that are identical. The design involved orthogonally crossing the factors: sample size (levels will be presented later), within-group distributions (normal distribution, normal distribution with outliers, and log-normal distribution), outlier removal method (z-score, inter quartile range [IQR], and median absolute deviation [MAD]), and cutoffs (1.5, 2, or 3 times the z-score/IQR distance/MAD). All the outlier removal methods remove a data point if it is considered too far from the other data. They only differ in the distance measure they use. The z-score approach uses a data point's distance from the mean, expressed in units of standard deviation; the MAD approach uses a data point's distance from the median, expressed in units of the MAD; and the IQR approach uses a data point's distance from the upper and lower quartile of the distribution, expressed in units of the IQR. The cutoffs define the distance at which a data point is considered an outlier. All of the outlier removal approaches were applied across groups. The investigated outlier removal methods and cutoffs closely mirror how outliers are removed in practice. This was examined by Bakker and Wicherts (2014), whose findings I summarize here. The z-score method with cutoffs of 2 and 3 is the most commonly used approach. While the IQR and MAD methods are rarely used in practice, they are considered to be improvements over the z-score approach. I therefore included them to demonstrate that removing outliers across groups is also invalid for these supposedly superior approaches.

I will now describe the aspects that differ from André (2022). The crucial difference was that I generated data such that the null hypothesis of no differences was false. In the first step, I generated data for each group in the same way as André: I drew data from one of the three within-group distributions for both groups and then randomly distributed them into two groups, so that there were no differences between the groups. However, I added 0.8 to the second group in a second step.[1] I chose 0.8 as it corresponds to Cohen's d of 0.8 for the normal distribution condition and thus to a large effect size, according to commonly used heuristics (Cohen, 1988). For didactical reasons, I will refer to the second group as the treatment group and the first as the control group. However, the arguments presented here also apply to any other situation in which two groups are compared.

The second main difference was that I added variance ratio (VR) as an additional factor with the levels 1 and 4. For $VR = 1$, the within-group variances were equal, as was the case in André (2022), whereas for $VR = 4$, the variance of the treatment group was 4 times larger than the variance of the control group. I chose 4 because this was the average VR found in empirical data (Erceg-Hurn and Mirosevich, 2008).

The first minor difference was in the statistical tests. Like André (2022), I used tests to establish whether the values of a variable of interest tend to be larger in one group, as this is the most pervasive research question in (experimental) psychology, and included the standard parametric test: Welch's t-test.[2] However, instead of the Wilcoxon–Mann–Whitney test, the standard nonparametric test, I included Brunner–Munzel's test, because it is the improved, recommended version of the Wilcoxon–Mann–Whitney test (Karch, 2021b; Noguchi et al., 2021). The second minor difference was that in addition to the sample sizes considered by André (50, 100, and 250), I also added the sample sizes 1, 000 and 10, 000. This made it possible to examine the large sample properties of removing outliers across groups.

As outcomes, I obtained the relative biases[3] of the parameter estimators, the coverage probabilities of the 95% confidence intervals returned by the tests, and the power of the tests. As well as raw power values, I report the difference in power compared to not removing outliers. If this power difference was positive, removing outliers increased the power. I estimated all quantities by generating random data for each design cell 100, 000 times. I also used 100, 000 replications for the other simulation studies. To ensure reproducibility, I share the code used for running this and all other simulation studies at https://osf.io/xt8zd/.

[1] In the supplemental material (https://osf.io/xt8zd/), I also present results for a smaller mean difference (0.4). However, since the results are qualitatively the same, I focus on 0.8 in this paper.

[2] Despite the nonnormally distributed data, applying Welch's t-test produces accurate results (p-values, parameter estimates, and confidence intervals), due to the relatively large sample sizes (e.g., Delacre et al., 2017; Field, 2017, Chapter 6), and it is therefore an appropriate analysis choice.

[3] Relative bias is defined as bias divided by true value. A relative bias of 100% thus means that, on average, a parameter estimate is 100% larger than the true value.

Results: Removing Outliers Across Groups Invalidates Difference Estimates and Confidence Intervals

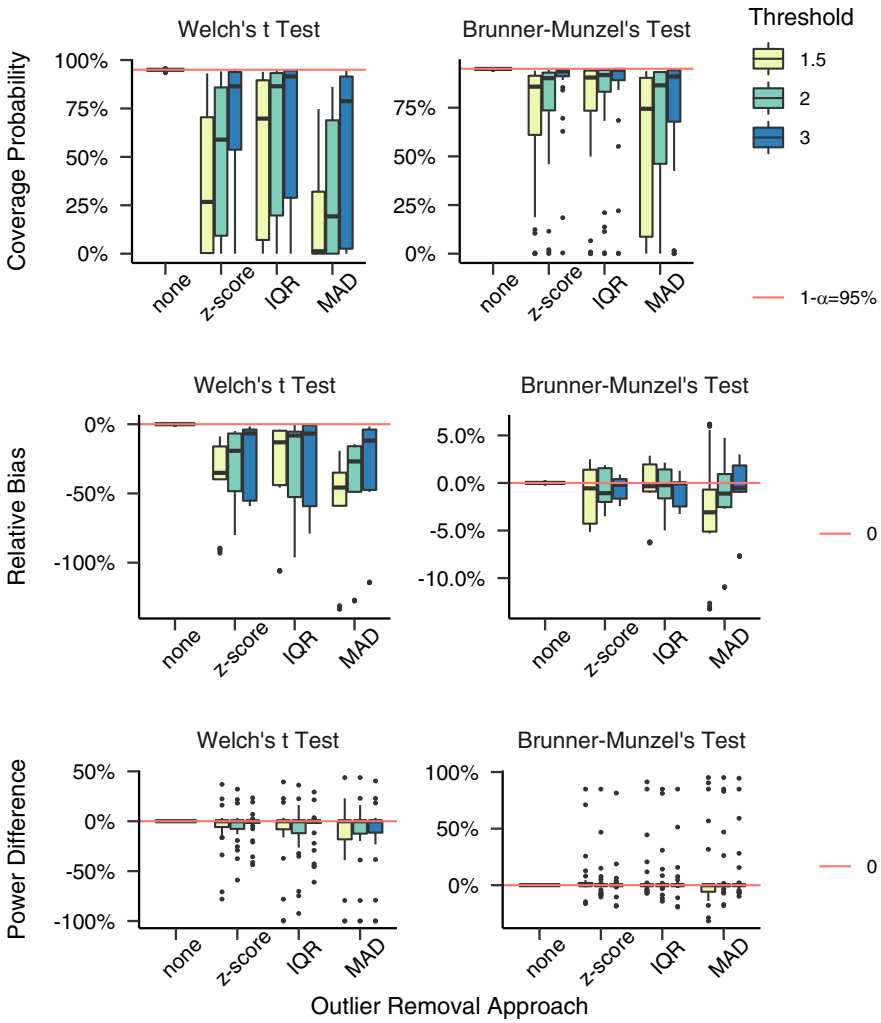
The full results are shown in Appendix A. In Figure 1, I summarize the results for each outlier removal method and cutoff combination, collapsed across the other factors. Without removing outliers, the coverage probabilities were always close to the nominal coverage probability of 95% (93%–96%). In contrast, when removing outliers, the coverage probabilities were typically much reduced. Without removing outliers, the estimates were never biased. All the estimated relative biases were between −0.30% and 0.09%. Mathematical derivations indeed show that the true bias is 0 for both Welch's t (Lehmann and Casella, 2003, Example 4.7) and Brunner–Munzel's test (Brunner et al., 2019, Result 3.1) for all the simulated conditions. In contrast, when removing outliers across groups, the estimates were typically biased.

The problems were evident across all the sample sizes, tests, outlier removal techniques, distributions, cutoffs, and VRs (see Appendix A). With larger cutoffs, the problems became less pronounced: coverage probabilities increased and biases decreased. However, problems were still severe even for the largest cutoff (3): coverage probabilities were as low as 0%. In some conditions, the confidence intervals thus never contained the true value. Relative bias was as great as −115% and hence substantial enough to wrongly conclude that the values in the control group were larger.

Problems were found across all sample sizes. Bias did not depend on the sample size. For example, for the condition with distribution "log-normal," outlier removal method "MAD," cutoff 2, VR = 1, and Welch's t-test, the estimated relative bias was −16%, regardless of the sample size. The coverage reduced as the sample size increased; for the condition just mentioned, for example, the coverage values steadily decreased from 86% for the smallest sample size to 0% for the largest sample size.

Figure 1

Summary of Results When Removing Outliers Across Groups



Note. The red lines mark the nominal coverage probability of 100% − α = 95%, a bias of 0, and a power difference of 0, respectively. See the online article for the color version of this figure.

In terms of power, the results are more ambiguous. Removing outliers across groups can increase or decrease the power. In some cases, the power was even improved when the differences were underestimated, which might seem paradoxical. I will discuss this result in more detail in the next section.

Explanation: Removing Outliers Across Groups Biases Difference Estimates and Standard Errors

To explain why removing outliers across groups leads to the presented results, I will focus on a few selected conditions. First, I will look at the condition: distribution "normal," sample size 250, outlier removal method "*z*-score," cutoff 2, VR = 1, and Welch's *t*-test. The relative bias was −18% and the coverage probability was 56% for this condition. Figure 2 shows why: without removing outliers, the estimated mean difference is close to the true mean difference, and the confidence interval contains the true value. In the control group, primarily values below the mean (of the control group) are removed, whereas, in the treatment group, primarily values above the mean (of the treatment group) are removed. This is caused by the difference in means between the two groups. As a consequence, after outlier removal, the sample mean of the control group is artificially increased and the sample mean of the treatment group is artificially decreased. This, in turn, leads to an underestimation of the true mean difference and the outcome that the confidence interval does not contain the true value.

The fact that the bias induced by removing outliers is constant across sample sizes, as discussed earlier, explains why the coverage steadily decreases with increasing sample size. As the sample size increases, the standard error and hence the width of the confidence interval decrease. Consequently, the confidence interval contains only a narrow range of values around the biased estimated mean difference in large samples.

It should be noted that although removing outliers always leads to an underestimation of the mean difference for all conditions examined in the simulation study, it can also lead to overestimating the difference. This is the case, for example, for the following scenario, which was not included in the design of the simulation study: distribution "log-normal," outlier removal method "MAD," cutoff 2, the true mean difference in favor of the treatment group 0.2, and control group having four times more variance. In this scenario, removing outliers across groups overestimates the mean difference by 191%.

As well as biasing parameter estimates, removing outliers across groups can also lead to underestimating the standard error. This explains the seemingly paradoxical result that the power can be improved by removing outliers, despite an underestimated mean difference. I will explain this on the basis of the condition: distribution "log-normal," sample size 50, outlier removal method "MAD," cutoff 3, VR = 1, and Welch's *t*-test. For this condition, the mean difference was underestimated by 11%. However, the power was greater (94.90%) than without removing outliers (54.52%). Figure 3 shows why: without removing outliers, the confidence interval is relatively wide and contains 0. Consequently, the null hypothesis is not rejected. In both groups, the largest values are removed, which leads to a decrease in variance for both groups, giving a reduced standard error and hence a narrower confidence interval. As a result, the confidence interval does not contain 0 and the null hypothesis is rejected. The fact that removing outliers increases the power in some conditions might seem like a favorable property. However, it is not, as it is caused by underestimating the uncertainty in the parameter estimates, which invalidates confidence intervals.

In summary, if the values in one group tend to be larger, removing outliers biases the parameter and uncertainty estimates, and also invalidates confidence intervals.

Null Hypothesis True: Removing Outliers Across Groups Can Bias Difference Estimates and Inflate Type I Error Rates

In this section, I will demonstrate that removing outliers can also lead to problems if the values in both groups tend to be equally large; that is, if the null hypothesis is true.

Design

The design for this simulation study was almost identical to the design used for the first one. Specifically, I used the same sample sizes, within-group distributions, outlier removal methods, cutoffs, and statistical tests. The first difference was that the data were created such that the null hypotheses of the tests were true; that is, I did not add 0.8 to the treatment group. The second difference was that I only considered a VR of VR = 4. I did not consider VR = 1 here because this would give completely equal distributions across groups, and removing outliers across groups does not invalidate statistical results for this scenario, as André (2022) demonstrated.

I obtained Type I error rates and biases[4] as outcome variables.

Results: Removing Outliers Across Groups Biases Difference Estimates and Inflates Type I Error Rates

The full results are shown in Appendix B. Figure 4 summarizes the results for each outlier removal method and cutoff combination, collapsed across the other factors. Without removing outliers, the Type I error rates were always close to the nominal error rate of 5% (5%–7%). In contrast, the Type I error rate was often severely inflated when removing outliers (5%–100%). Similarly, without removing outliers, the parameter estimates were never biased. In contrast, when removing outliers across groups, the difference between the groups was typically overestimated (bias range: 0.04–0.85).

The problems were evident in almost all conditions. The only exception was normal data, for which no bias was observed and the type I error rates were only slightly inflated. The bias was again constant across sample sizes (see Table B2), and the Type I error rate became greater with increasing sample sizes (see Table B1).

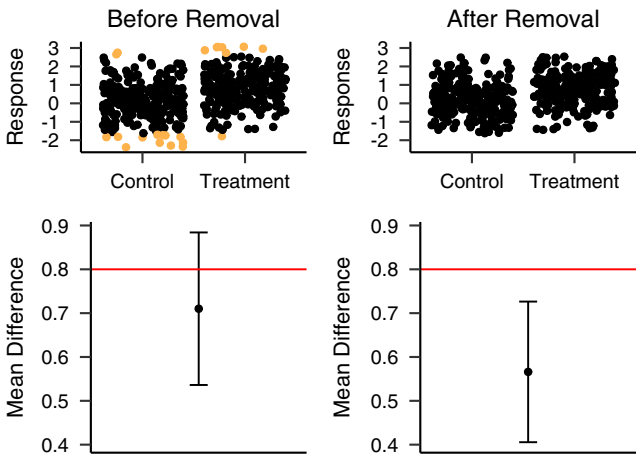
Explanation: Removing Outliers Across Groups Creates False Group Differences

To explain the results, I will focus on the following condition: distribution "log-normal," sample size 250, outlier removal method "*z*-score," cutoff 3, and Welch's *t*-test. Note that the Type I error rate was 60.69% for this condition. Figure 5 visualizes

[4] I chose bias here instead of relative bias, because the relative bias is not defined when the true mean difference is 0.

Figure 2

Demonstration That Removing Outliers Across Groups Introduces Bias

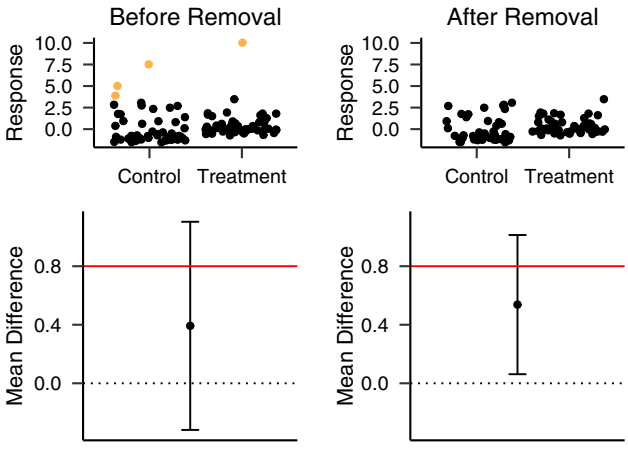


Note. Responses that are considered outliers are marked orange. The uncertainty bands visualize the 95% confidence intervals for the difference in means, as obtained by Welch's *t*-test. The red line marks the true mean difference. See the online article for the color version of this figure.

why: without removing outliers, the sample means are almost identical and Welch's *t*-test correctly does not reject the null hypothesis of equal means. Due to the positive skew of the data, only large values (> 8.31) are removed in both groups. Consequently, in both groups, removing outliers reduces the mean. However, in the treatment group, many more values exceed the cutoff of 8.31, because it has a larger variance. As a result, the mean is reduced more in the treatment group, which creates a false

Figure 3

Demonstration That Removing Outliers Leads to Underestimated Standard Errors



Note. Responses that are considered outliers are marked orange. The uncertainty bands visualize the 95% confidence intervals for the difference in means, as obtained by Welch's *t*-test. The red line marks the true mean difference. The dashed line marks 0, the mean difference postulated by the null hypothesis. See the online article for the color version of this figure.

mean difference. At the same time, the variance in both groups and, hence the width of the confidence intervals are substantially decreased. In combination, these two phenomena lead to incorrectly rejecting the null hypothesis.

In summary, if the distributions of the two groups are different, for example, due to unequal variances, there is a substantial probability that removing outliers across groups will lead to falsely concluding that the values in one group tend to be larger.

Null Hypothesis False: Hypothesis-blind Outlier Exclusion Invalidates Regression Estimates and Confidence Intervals

In this last section, I will show that also for regression models, hypothesis-blind outlier removal invalidates confidence intervals and biases parameter estimates if the null hypothesis is false. As a reminder: removing outliers across groups is a special case of the more general strategy of hypothesis-blind outlier removal. For regression, hypothesis-blind outlier removal involves removing outliers using a regression model that does not contain the independent variables of interest. More specifically, consider a regression model that tests whether an independent variable X is associated with a dependent variable Y, controlling for a covariate W. The regression model testing this hypothesis is

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 W_i + \epsilon_i \quad (1)$$

where ϵ_i represents the residuals and the null hypothesis is $H_0 : \beta_1 = 0$. A common outlier removal approach is to investigate the model residuals, that is, the difference between the predicted and actual values of the dependent variable, and to remove data points with residuals exceeding a certain cutoff. Because it uses the full model shown in Equation 1 to calculate the residuals, this approach is hypothesis-aware (André, 2022). In contrast, the hypothesis-blind approach uses the following reduced model to obtain the residuals

$$Y_i = \beta_0 + \beta_2 W_i + \epsilon_i. \quad (2)$$

This crucially does not include the independent variable X as a predictor.

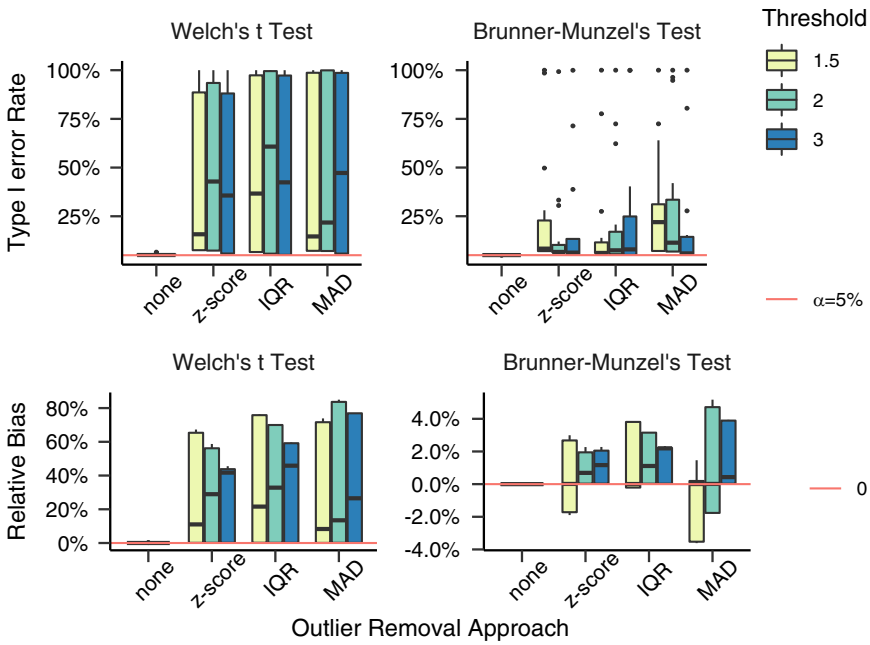
Design

The design was again almost the same as the one used by André (2022). More specifically, the full model shown in Equation 1 was used to generate the data. Residuals were obtained from the reduced model (Equation 2). Data points with a standardized residual < -2 or > 2 were removed. The design involved orthogonally crossing the factors sample size (50, 100, 250, 1,000, and 10,000), distribution of the independent variable X (normal and categorical), and distribution of the error ϵ (normal, normal with outliers, and log-normal).[5] The covariate W was always distributed according to a standard normal distribution. As the true regression coefficients for the intercept and the covariate, 1 was

[5] See the code in the supplemental material (https://osf.io/xt8zd/) for a detailed description of the distributions used.

Figure 4

Type I Error Rates and Biases When Removing Outliers Across Groups

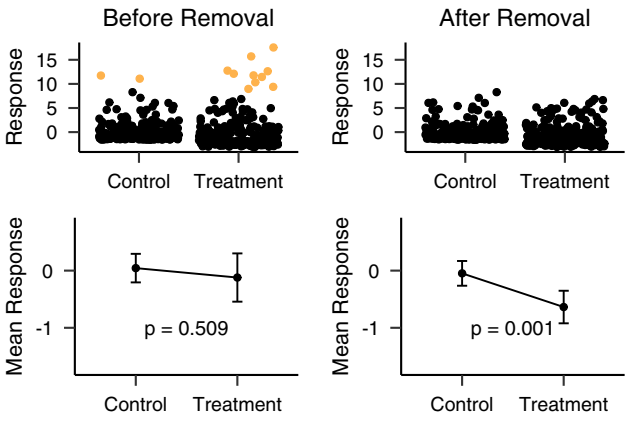


Note. The red lines mark the nominal type I error rate of $\alpha = 5\%$, and a bias of 0 respectively. See the online article for the color version of this figure.

used ($\beta_0 = \beta_1 = 1$). The only relevant difference compared to André is that they set the true coefficient for the variable of interest X at $\beta_1 = 0$, as they were focusing on Type I error rates for testing H_0:$\beta_1 = 0$, whereas I set the true coefficient at $\beta_1 = 1$, and focused on relative biases and the coverage probabilities for the coefficient β_1.

Figure 5

Demonstration That Removing Outliers Across Groups Creates False Group Differences



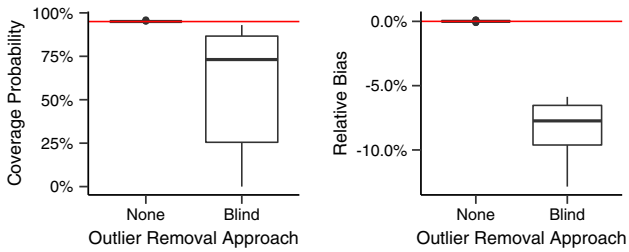
Note. Responses that are considered outliers are marked orange. The *p*-value is obtained using Welch's *t*-test. The uncertainty bands visualize the 95% confidence intervals for the group means. See the online article for the color version of this figure.

Results: Hypothesis-Blind Outlier Removal Invalidates Parameter Estimates and Confidence Intervals

The full results are shown in Appendix C. In Figure 6, I summarize the results. The coverage probabilities were always reduced for hypothesis-blind outlier removal (0%–93%). In contrast, they were always close to the nominal value of 95% without outlier removal (95%–96%). Similarly, when removing outliers, the estimates were always substantially biased (relative bias range: −13% to −6%), whereas this was never the case without removing outliers.

Figure 6

Coverage Probabilities and Relative Biases For Hypothesis-Blind Outlier Removal in a Regression Model



Note. The boxplots visualize the coverage probabilities and relative biases obtained across all cells of the simulation study, that is, across different sample sizes and distributions for the dependent and error variables. The red lines mark the nominal coverage probability of $100\% - \alpha = 95\%$, and a bias of 0, respectively. See the online article for the color version of this figure.

Discussion

André (2022) recommended hypothesis-blind outlier removal instead of hypothesis-aware outlier removal, arguing that it does not lead to inflated Type I error rates. In contrast, I demonstrated that hypothesis-blind outlier removal leads to problems similar to those created by hypothesis-aware outlier removal. Specifically, when the null hypothesis is false, it is almost always the case that parameter estimates are biased and confidence intervals invalidated. Problems are less severe when the null hypothesis is true. However, when comparing two groups, inflated Type I error rates and biased parameter estimates still occur if variances are unequal and data nonnormal.

The primary reason that the disadvantages of hypothesis-blind outlier removal were not revealed by André (2022) was their focus on Type I error rates. As they explained, hypothesis-aware outlier removal assumes that the null hypothesis is false, which leads to problems if the null hypothesis is true. However, hypothesis-blind outlier removal assumes that the null hypothesis is true, which leads to problems if it is false. This was not uncovered by André, as they only investigated situations where the null hypothesis *was true*. In contrast, I showed that when the null hypothesis *is false*, hypothesis-blind outlier removal almost always invalidates statistical results.

André did not uncover that for comparing groups hypothesis-blind outlier removal can also lead to problems when the null hypothesis *is true* for the following reasons. Removing outliers across groups not only assumes that the null hypothesis is true but also that the distribution of values is identical in all aspects across the groups. If this is not the case, for example, due to variance differences, removing outliers across groups can inflate Type I error rates and bias parameter estimates, as I demonstrated. This disadvantage of hypothesis-blind outlier removal was not revealed by *André* because, in their simulation studies, the distributions were always identical across groups

In practice, hypothesis-blind outlier removal is more problematic when the null hypothesis *is false*. If the null hypothesis is false, the central assumption of hypothesis-blind outlier removal (null hypothesis is true) is violated. Consequently, hypothesis-blind outlier removal essentially guarantees that parameter estimates and confidence intervals are invalidated (see Figures 1 and 6). This is the main reason to avoid hypothesis-blind outlier removal.

How often hypothesis-blind outlier removal will lead to problems in practice when the null hypothesis *is true* is less certain. It depends on whether it is likely that the values of two groups will tend to be equally large but differ in some other property, such as having unequal variances, and deviating from normality. Concerning normality, the evidence is clear: real data are seldom perfectly normal (Micceri, 1989). With regard to unequal variances, Grissom (2000) claimed that it is "difficult to find an example in which … groups differ in variability but not in means." However, there are examples where this was the case (Dykiert et al., 2012; Johnson et al., 2008). Whether one can assume equal variances under the null or not is prominently discussed in the literature concerning the proper usage of the *t*-test. The consensus is not to assume equal variances and instead use the unequal variances *t*-test by default (Delacre et al., 2017; Hayes and Cai, 2007; Ruxton, 2006). The reasoning is that for a particular data set, it is almost impossible to determine whether population variances are equal; Type I error rates can be inflated if variances are unequal; and not much is lost by giving up the equal variances assumption (Delacre et al., 2017). Similarly, for hypothesis-blind outlier removal, it is almost impossible to exclude the possibility that the groups differ in some property that does not lead to the values in one group tending to be larger; Type I error rates can be inflated if this is the case; and not much is lost by giving up this assumption (see the next section). Thus, even considering Type I error rates alone, outliers should not be removed using a hypothesis-blind procedure.

Synthesizing the arguments presented by André (2022) and this paper thus leads to the following conclusion. Data points must not be automatically removed solely because they are deemed outliers, whether the procedure used is hypothesis-aware or hypothesis-blind.

How to Handle Outliers Correctly

In this subsection, I advise on how to handle outliers correctly. The general principle of my advice is not to remove outliers but instead to use robust statistical methods. This has already been recommended as a valid alternative to *hypothesis-aware* outlier removal (Bakker and Wicherts, 2014; Wilcox, 1998), and André (2022) also briefly discussed robust techniques as a valid alternative to *hypothesis-blind* outlier removal.

First, outliers and erroneous values should be treated differently (Stevens, 2012). As a reminder: an outlier is a value that is too far removed from the rest of the data, whereas an erroneous value can clearly be identified as incorrect, for example, due to malfunction of a measurement device, errors during data input, or just being deemed impossible. Importantly, a data point can be erroneous but not an outlier, and vice versa.

Erroneous values should be identified and then corrected or deleted (Stevens, 2012). After removing incorrect values, one can run a standard technique. However, this implicitly assumes that there are no—or very few—erroneous values left, as even a small amount of erroneous data can completely invalidate the results of standard techniques, especially if they are also outliers (Wilcox, 2016). Using standard techniques is thus only appropriate if it is certain that the cleaned data contain almost no erroneous values.

Otherwise, robust approaches should be employed. The central idea of robust statistics is that instead of safeguarding the sensitive standard techniques by removing outliers, outliers are not removed; rather, techniques are designed to be inherently robust to outliers (Wilcox, 2016). For comparing two means, the robust approach that most closely implements the general philosophy of removing outliers is Yuen–Welch's test, as it completely ignores the most extreme values within each group. It is thus akin to hypothesis-aware outlier removal. However, it crucially does not suffer from Type I error inflation, biased parameter estimates, or invalid confidence intervals (Wilcox, 1998). Hence, it can be recommended as the default alternative to removing outliers when comparing two means. There are, nevertheless, many viable alternatives, such as comparing M-estimators and medians. I refer the reader to Wilcox (2016), Chapters 8 and 12 for more extensive guidance in choosing the appropriate approach.

Nonparametric, rank-based tests for comparing two groups, like the Wilcoxon–Mann–Whitney or the better Brunner–Munzel test, are already reasonably robust to outliers, because they only consider the rank order of the data. Consequently, these tests can be recommended for use without removing outliers. It is important to note,

however, that they do not test the equality of means (Karch, 2021a) and are therefore not appropriate alternatives for comparing means.

For regression, a multitude of robust techniques exist. Wilcox (2016), Chapter 15 provides an overview that will help in selecting the most appropriate technique for a particular task.

Transparency and Openness

The OSF repository of the project (https://osf.io/xt8zd/) contains supplemental material as well as the code and files needed to reproduce all the analyses, the figures, and the code for typesetting the article itself.

References

André, Q. (2022). Outlier exclusion procedures must be blind to the researcher's hypothesis. *Journal of Experimental Psychology: General*, *151*(1), 213–223. https://doi.org/10.1037/xge0001069

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. https://doi.org/10.1037/met0000014

Brunner, E., Bathke, A. C., & Konietschke, F. (2019). *Rank and pseudo-rank procedures for independent observations in factorial designs: Using R and SAS*. Springer. https://doi.org/10.1007/978-3-030-02914-2

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum Associates.

Delacre, M., Lakens, D., & Leys, C. (2017). Why psychologists should by default use Welch's *t*-test instead of Student's *t*-test. *International Review of Social Psychology*, *30*(1), 92–101. https://doi.org/10.5334/irsp.82

Dykiert, D., Der, G., Starr, J. M., & Deary, I. J. (2012). Sex differences in reaction time mean and intraindividual variability across the life span. *Developmental Psychology*, *48*(5), 1262–1276. https://doi.org/10.1037/a0027550

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

Field, A. (2017). *Discovering statistics using IBM SPSS statistics* (5th ed.). Sage.

Grissom, R. J. (2000). Heterogeneity of variance in clinical data. *Journal of Consulting and Clinical Psychology*, *68*(1), 155–165. https://doi.org/10.1037/0022-006X.68.1.155

Hayes, A. F., & Cai, L. (2007). Further evaluating the conditional decision rule for comparing two independent means. *British Journal of Mathematical and Statistical Psychology*, *60*(2), 217–244. https://doi.org/10.1348/000711005X62576

Howitt, D., & Cramer, D. (2007). *Introduction to statistics in psychology*. Pearson Education.

Johnson, W., Carothers, A., & Deary, I. J. (2008). Sex differences in variability in general intelligence: A new look at the old question. *Perspectives on Psychological Science*, *3*(6), 518–531. https://doi.org/10.1111/j.1745-6924.2008.00096.x

Karch, J. D. (2021a). *Choosing between the two-sample* t *test and its alternatives: A practical guideline*. https://doi.org/10.31234/osf.io/ye2d4

Karch, J. D. (2021b). Psychologists should use Brunner–Munzel's instead of Mann–Whitney's *U* test as the default nonparametric procedure. *Advances in Methods and Practices in Psychological Science*, *4*(2), Article 251524592199960. https://doi.org/10.1177/2515245921999602

Lehmann, E. L., & Casella, G. (2003). *Theory of point estimation* (2nd ed. 1998, Corr. 4th printing). Springer.

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin*, *105*(1), 156–166. https://doi.org/10.1037/0033-2909.105.1.156

Noguchi, K., Konietschke, F., Marmolejo-Ramos, F., & Pauly, M. (2021). Permutation tests are robust and powerful at 0.5% and 5% significance levels. *Behavior Research Methods*, *53* (6), 2712–2724. https://doi.org/10.3758/s13428-021-01595-5

Pincus, R., & Lewis, T. (1994). *Outliers in statistical data*. Wiley.

Ruxton, G. D. (2006). The unequal variance *t*-test is an underused alternative to Student's *t*-test and the Mann–Whitney *U* test. *Behavioral Ecology*, *17*(4), 688–690. https://doi.org/10.1093/beheco/ark016

Stevens, J. P. (2012). *Applied multivariate statistics for the social sciences*. Routledge.

Wilcox, R. R. (1998). How many discoveries have been lost by ignoring modern statistical methods? *American Psychologist*, *53*(3), 300–314. https://doi.org/10.1037/0003-066X.53.3.300

Wilcox, R. R. (2016). *Introduction to robust estimation and hypothesis testing* (4th ed.). Elsevier.

(*Appendices follow*)

Appendix A

Results of Comparing Groups: Null Hypothesis False

Equal Variances

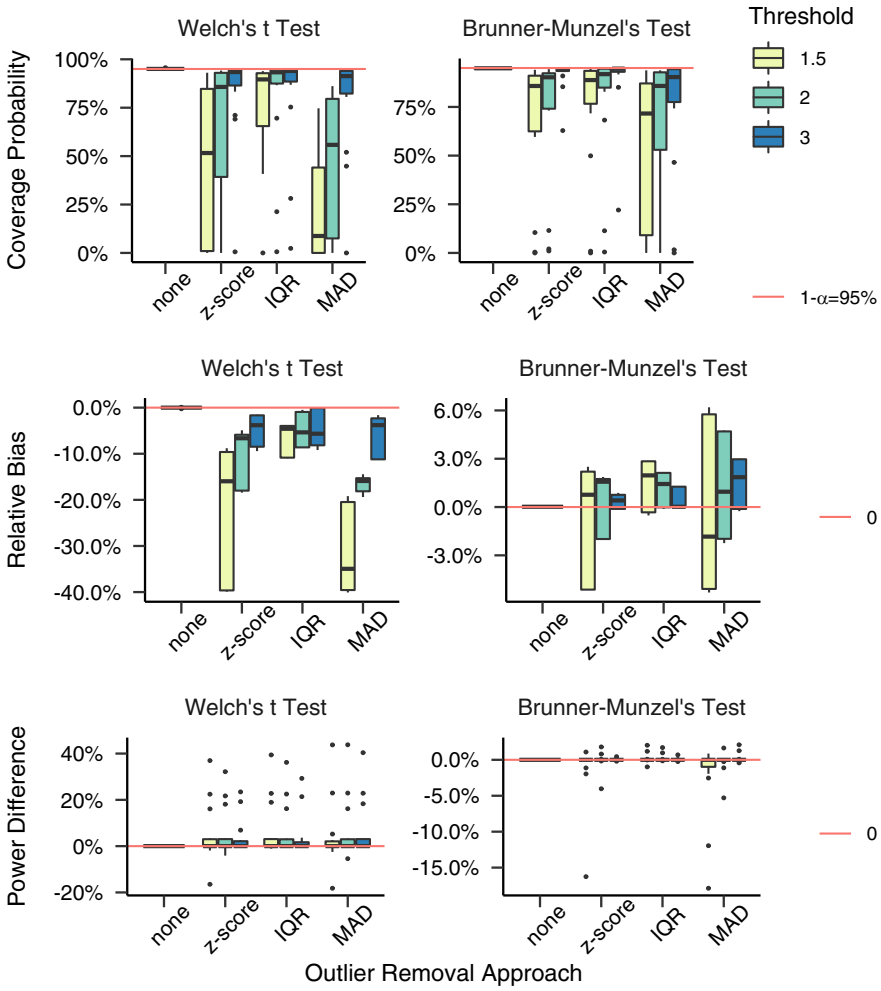
See Figure A1 and Tables A1, A2, and A3.

Unequal Variances

See Figure A2 and Tables A4, A5, and A6.

Figure A1

Summary of Results When Removing Outliers Across Groups; Equal Variances



Note. The red lines mark the nominal coverage probability of 100% − α = 95%, a bias of 0, and a power difference of 0 respectively. See the online article for the color version of this figure.

(*Appendices continue*)

Table A1
Coverage Probabilities When Removing Outliers Across Groups; Equal Variances

				IQR			MAD			*z*-score		
Test	Data	*N*	None	1.5	2	3	1.5	2	3	1.5	2	3
Welch	Normal	50	.95	**.93**	.95	.95	**.49**	**.82**	.94	**.52**	**.86**	.95
Welch	Normal	100	.95	**.93**	.95	.95	**.26**	**.75**	.94	**.24**	**.78**	.94
Welch	Normal	250	.95	**.92**	.95	.95	**.03**	**.55**	**.94**	**.01**	**.56**	.94
Welch	Normal	1,000	.95	**.88**	.95	.95	**.00**	**.09**	**.93**	**.00**	**.06**	**.93**
Welch	Normal	10,000	.95	**.41**	**.94**	.95	**.00**	**.00**	**.84**	**.00**	**.00**	**.83**
Welch	Normal mixture	50	.95	**.94**	.95	.94	**.60**	**.86**	.95	**.87**	.94	.95
Welch	Normal mixture	100	.95	**.94**	.95	**.93**	**.39**	**.82**	.95	**.82**	**.94**	**.93**
Welch	Normal mixture	250	.95	**.93**	**.93**	**.90**	**.09**	**.69**	**.94**	**.67**	**.92**	**.90**
Welch	Normal mixture	1,000	.95	**.89**	**.87**	**.75**	**.00**	**.25**	**.91**	**.17**	**.86**	**.71**
Welch	Normal mixture	10,000	.95	**.43**	**.21**	**.02**	**.00**	**.00**	**.52**	**.00**	**.22**	**.01**
Welch	Log-normal	50	.96	**.92**	**.93**	.94	**.75**	**.86**	**.91**	**.93**	.94	.95
Welch	Log-normal	100	.95	**.90**	**.92**	**.94**	**.63**	**.78**	**.89**	**.92**	**.94**	.95
Welch	Log-normal	250	.95	**.82**	**.88**	**.93**	**.32**	**.56**	**.81**	**.88**	**.92**	.94
Welch	Log-normal	1,000	.95	**.49**	**.70**	**.87**	**.00**	**.06**	**.45**	**.67**	**.83**	**.92**
Welch	Log-normal	10,000	.95	**.00**	**.01**	**.28**	**.00**	**.00**	**.00**	**.00**	**.15**	**.69**
BM	Normal	50	.95	.94	.95	.95	**.90**	**.94**	.95	**.91**	.94	.95
BM	Normal	100	.95	.95	.95	.95	**.84**	**.93**	.95	**.85**	**.94**	.95
BM	Normal	250	.95	.95	.95	.95	**.65**	**.90**	.95	**.65**	**.90**	.95
BM	Normal	1,000	.95	.94	.95	.95	**.12**	**.73**	.95	**.10**	**.73**	.95
BM	Normal	10,000	.95	**.89**	.95	.95	**.00**	**.01**	.94	**.00**	**.01**	.94
BM	Normal mixture	50	.95	**.93**	**.94**	.95	**.94**	**.93**	**.93**	**.94**	**.93**	.94
BM	Normal mixture	100	.95	**.92**	**.93**	.95	**.93**	**.93**	**.92**	.94	**.92**	.94
BM	Normal mixture	250	.95	**.89**	**.92**	.95	**.91**	**.93**	**.90**	**.94**	**.90**	.94
BM	Normal mixture	1,000	.95	**.72**	**.83**	.95	**.80**	**.89**	**.74**	**.91**	**.75**	**.94**
BM	Normal mixture	10,000	.95	**.01**	**.11**	.95	**.07**	**.44**	**.02**	**.60**	**.02**	**.85**
BM	Log-normal	50	.95	**.90**	**.92**	**.93**	**.81**	**.86**	**.90**	**.91**	**.92**	**.94**
BM	Log-normal	100	.95	**.88**	**.91**	**.93**	**.72**	**.80**	**.88**	**.90**	**.92**	**.94**
BM	Log-normal	250	.95	**.82**	**.87**	**.92**	**.45**	**.62**	**.81**	**.86**	**.90**	**.93**
BM	Log-normal	1,000	.95	**.50**	**.68**	**.85**	**.02**	**.11**	**.47**	**.66**	**.81**	**.91**
BM	Log-normal	10,000	.95	**.00**	**.00**	**.22**	**.00**	**.00**	**.00**	**.00**	**.12**	**.63**

Note. Coverage probabilities are boldface if they are substantially too low (< 0.94). N = sample size, BM = Brunner–Munzel's test. None, IQR, MAD, and *z*-score were the outlier removal strategies. IQR = interquartile range; MAD = median absolute deviation. The numbers (1.5, 2, 3) refer to the cutoff used by the respective outlier removal approach.

(*Appendices continue*)

Table A2
Relative Biases When Removing Outliers Across Groups; Equal Variances

				IQR			MAD			*z*-score		
Test	Data	*N*	None	1.5	2	3	1.5	2	3	1.5	2	3
Welch	Normal	50	0.00	**−0.05**	**−0.01**	−0.00	**−0.40**	**−0.19**	**−0.03**	**−0.40**	**−0.18**	**−0.02**
Welch	Normal	100	0.00	**−0.04**	−0.01	−0.00	**−0.40**	**−0.19**	**−0.02**	**−0.40**	**−0.18**	**−0.02**
Welch	Normal	250	0.00	**−0.04**	−0.01	−0.00	**−0.40**	**−0.18**	**−0.02**	**−0.40**	**−0.18**	**−0.02**
Welch	Normal	1,000	0.00	**−0.04**	−0.00	0.00	**−0.40**	**−0.18**	**−0.02**	**−0.40**	**−0.18**	**−0.02**
Welch	Normal	10,000	0.00	**−0.04**	−0.00	−0.00	**−0.40**	**−0.18**	**−0.02**	**−0.40**	**−0.18**	**−0.02**
Welch	Normal mixture	50	0.00	**−0.05**	**−0.06**	**−0.08**	**−0.35**	**−0.16**	**−0.05**	**−0.17**	**−0.06**	**−0.08**
Welch	Normal mixture	100	0.00	**−0.05**	**−0.06**	**−0.08**	**−0.35**	**−0.15**	**−0.04**	**−0.16**	**−0.06**	**−0.09**
Welch	Normal mixture	250	0.00	**−0.04**	**−0.05**	**−0.09**	**−0.35**	**−0.15**	**−0.04**	**−0.16**	**−0.05**	**−0.09**
Welch	Normal mixture	1,000	0.00	**−0.04**	**−0.05**	**−0.09**	**−0.35**	**−0.15**	**−0.04**	**−0.16**	**−0.05**	**−0.09**
Welch	Normal mixture	10,000	0.00	**−0.04**	**−0.05**	**−0.09**	**−0.35**	**−0.14**	**−0.04**	**−0.16**	**−0.05**	**−0.09**
Welch	Log-normal	50	0.00	**−0.11**	**−0.09**	**−0.06**	**−0.21**	**−0.16**	**−0.11**	**−0.10**	**−0.07**	**−0.04**
Welch	Log-normal	100	0.00	**−0.11**	**−0.09**	**−0.06**	**−0.20**	**−0.16**	**−0.11**	**−0.09**	**−0.07**	**−0.04**
Welch	Log-normal	250	0.00	**−0.11**	**−0.09**	**−0.06**	**−0.19**	**−0.16**	**−0.11**	**−0.09**	**−0.07**	**−0.04**
Welch	Log-normal	1,000	0.00	**−0.11**	**−0.09**	**−0.06**	**−0.19**	**−0.16**	**−0.11**	**−0.09**	**−0.06**	**−0.04**
Welch	Log-normal	10,000	0.00	**−0.11**	**−0.09**	**−0.06**	**−0.19**	**−0.16**	**−0.11**	**−0.09**	**−0.06**	**−0.04**
BM	Normal	50	0.00	−0.01	−0.00	−0.00	**−0.05**	**−0.02**	−0.00	**−0.05**	**−0.02**	−0.00
BM	Normal	100	0.00	−0.00	−0.00	0.00	**−0.05**	**−0.02**	−0.00	**−0.05**	**−0.02**	−0.00
BM	Normal	250	0.00	−0.00	−0.00	0.00	**−0.05**	**−0.02**	−0.00	**−0.05**	**−0.02**	−0.00
BM	Normal	1,000	0.00	−0.00	−0.00	0.00	**−0.05**	**−0.02**	−0.00	**−0.05**	**−0.02**	−0.00
BM	Normal	10,000	0.00	−0.00	−0.00	0.00	**−0.05**	**−0.02**	−0.00	**−0.05**	**−0.02**	−0.00
BM	Normal mixture	50	0.00	**0.02**	**0.01**	0.00	**−0.02**	0.01	**0.02**	0.01	**0.02**	0.00
BM	Normal mixture	100	0.00	**0.02**	**0.01**	0.00	**−0.02**	0.01	**0.02**	0.01	**0.02**	0.00
BM	Normal mixture	250	0.00	**0.02**	**0.01**	−0.00	**−0.02**	0.01	**0.02**	0.01	**0.02**	0.00
BM	Normal mixture	1,000	0.00	**0.02**	**0.01**	−0.00	**−0.02**	0.01	**0.02**	0.01	**0.02**	0.00
BM	Normal mixture	10,000	0.00	**0.02**	**0.01**	−0.00	**−0.02**	**0.01**	**0.02**	0.01	**0.02**	0.00
BM	Log-normal	50	0.00	**0.03**	**0.02**	**0.01**	**0.06**	**0.05**	**0.03**	**0.03**	**0.02**	0.01
BM	Log-normal	100	0.00	**0.03**	**0.02**	**0.01**	**0.06**	**0.05**	**0.03**	**0.02**	**0.02**	0.01
BM	Log-normal	250	0.00	**0.03**	**0.02**	**0.01**	**0.06**	**0.05**	**0.03**	**0.02**	**0.02**	0.01
BM	Log-normal	1,000	0.00	**0.03**	**0.02**	**0.01**	**0.06**	**0.05**	**0.03**	**0.02**	**0.02**	0.01
BM	Log-normal	10,000	0.00	**0.03**	**0.02**	**0.01**	**0.06**	**0.05**	**0.03**	**0.02**	**0.01**	0.01

Note. Relative bias estimates are boldface if they are substantially different from zero (< 0.01 or > 0.01). $N =$ sample size; BM = Brunner–Munzel's test. None, IQR, MAD, and *z*-score were the outlier removal strategies. IQR = interquartile range; MAD = median absolute deviation. The numbers (1.5, 2, 3) refer to the cutoff used by the respective outlier removal approach.

(*Appendices continue*)

Table A3
Power When Removing Outliers Across Groups; Equal Variances

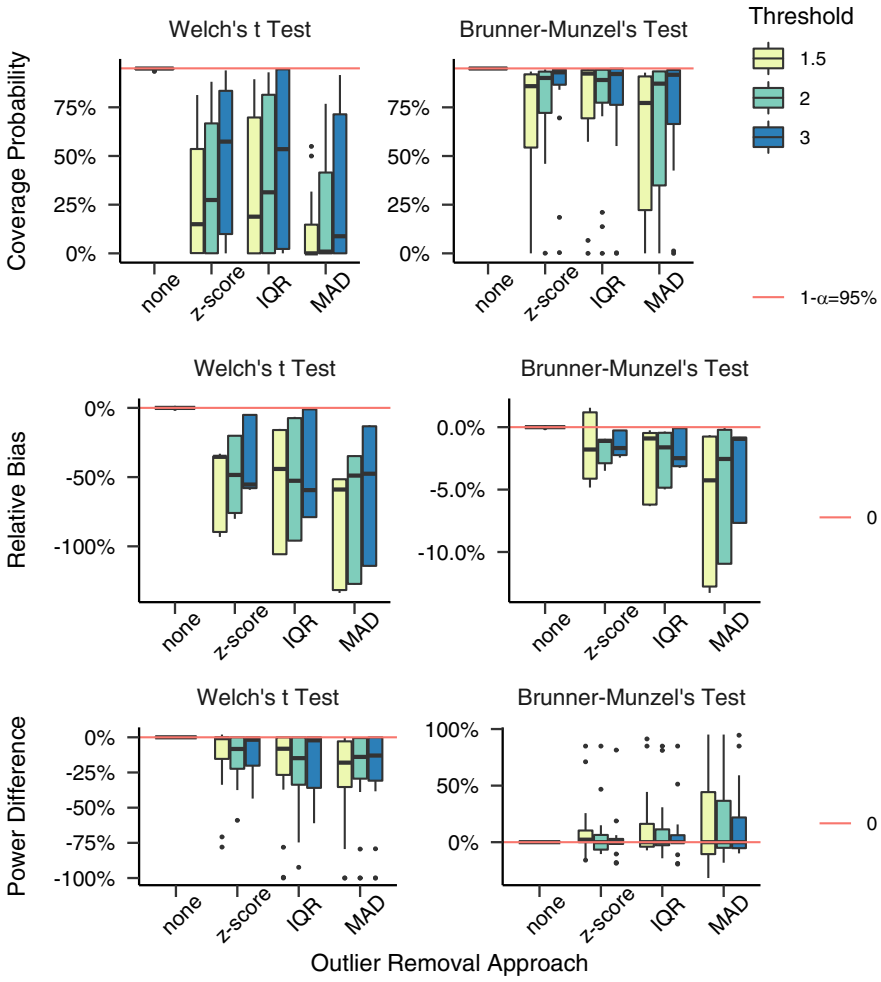
				IQR			MAD			*z*-score		
Test	Data	*N*	None	1.5	2	3	1.5	2	3	1.5	2	3
Welch	Normal	50	.98	**.97**	.98	.98	**.80**	**.92**	**.97**	**.81**	**.94**	.98
Welch	Normal	100	1.00	1.00	1.00	1.00	**.97**	1.00	1.00	**.98**	1.00	1.00
Welch	Normal	250	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Welch	Normal	1,000	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Welch	Normal	10,000	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Welch	Normal mixture	50	.77	.95	.93	.80	.82	.93	.95	.93	.95	.83
Welch	Normal mixture	100	.96	1.00	1.00	.97	.98	1.00	1.00	1.00	1.00	.98
Welch	Normal mixture	250	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Welch	Normal mixture	1,000	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Welch	Normal mixture	10,000	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Welch	Log-normal	50	.55	.94	.91	.84	.98	.98	.95	.91	.87	.78
Welch	Log-normal	100	.77	1.00	1.00	.98	1.00	1.00	1.00	.99	.99	.96
Welch	Log-normal	250	.97	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Welch	Log-normal	1,000	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Welch	Log-normal	10,000	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
BM	Normal	50	.97	**.96**	.97	.97	**.79**	**.92**	.97	**.81**	**.93**	.97
BM	Normal	100	1.00	1.00	1.00	1.00	**.97**	1.00	1.00	**.98**	1.00	1.00
BM	Normal	250	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
BM	Normal	1,000	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
BM	Normal	10,000	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
BM	Normal mixture	50	.93	.95	.95	.93	**.81**	**.92**	.96	**.92**	.95	.94
BM	Normal mixture	100	1.00	1.00	1.00	1.00	**.98**	1.00	1.00	1.00	1.00	1.00
BM	Normal mixture	250	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
BM	Normal mixture	1,000	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
BM	Normal mixture	10,000	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
BM	Log-normal	50	.98	.99	.99	.98	.98	.99	.99	.99	.98	.98
BM	Log-normal	100	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
BM	Log-normal	250	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
BM	Log-normal	1,000	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
BM	Log-normal	10,000	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

Note. Power estimates are boldface if removing outliers decreased the power. *N* = sample size; BM = Brunner–Munzel's test. None, IQR, MAD, and *z*-score were the outlier removal strategies. IQR = interquartile range; MAD = median absolute deviation. The numbers (1.5, 2, 3) refer to the cutoff used by the respective outlier removal approach.

(*Appendices continue*)

Figure A2

Summary of Results When Removing Outliers Across Groups; Unequal Variances



Note. The red lines mark the nominal coverage probability of $100\% - \alpha = 95\%$, a bias of 0, and a power difference of 0 respectively. See the online article for the color version of this figure.

(*Appendices continue*)

Table A4
Coverage Probabilities When Removing Outliers Across Groups; Unequal Variances

				IQR			MAD			*z*-score		
Test	Data	*N*	None	1.5	2	3	1.5	2	3	1.5	2	3
Welch	Normal	50	.95	**.89**	**.93**	.95	**.55**	**.77**	**.92**	**.76**	**.88**	**.94**
Welch	Normal	100	.95	**.87**	**.93**	.95	**.32**	**.65**	**.90**	**.62**	**.84**	**.94**
Welch	Normal	250	.95	**.81**	**.92**	.95	**.05**	**.36**	**.86**	**.29**	**.72**	**.93**
Welch	Normal	1,000	.95	**.52**	**.86**	.95	**.00**	**.01**	**.66**	**.00**	**.27**	**.90**
Welch	Normal	10,000	.95	**.00**	**.33**	.94	**.00**	**.00**	**.00**	**.00**	**.00**	**.52**
Welch	Normal mixture	50	.95	**.77**	**.77**	**.75**	**.50**	**.67**	**.77**	**.81**	**.78**	**.77**
Welch	Normal mixture	100	.95	**.62**	**.58**	**.55**	**.25**	**.47**	**.60**	**.71**	**.62**	**.57**
Welch	Normal mixture	250	.95	**.28**	**.19**	**.16**	**.02**	**.13**	**.24**	**.46**	**.26**	**.18**
Welch	Normal mixture	1,000	.95	**.00**	**.00**	**.00**	**.00**	**.00**	**.00**	**.02**	**.00**	**.00**
Welch	Normal mixture	10,000	.95	**.00**	**.00**	**.00**	**.00**	**.00**	**.00**	**.00**	**.00**	**.00**
Welch	Log-normal	50	**.93**	**.19**	**.31**	**.54**	**.00**	**.01**	**.09**	**.35**	**.51**	**.73**
Welch	Log-normal	100	**.94**	**.03**	**.10**	**.31**	**.00**	**.00**	**.01**	**.15**	**.32**	**.61**
Welch	Log-normal	250	.94	**.00**	**.00**	**.05**	**.00**	**.00**	**.00**	**.01**	**.07**	**.35**
Welch	Log-normal	1,000	.95	**.00**	**.00**	**.00**	**.00**	**.00**	**.00**	**.00**	**.00**	**.02**
Welch	Log-normal	10,000	.95	**.00**	**.00**	**.00**	**.00**	**.00**	**.00**	**.00**	**.00**	**.00**
BM	Normal	50	.95	**.94**	.94	.95	**.90**	**.92**	.94	**.93**	**.93**	.95
BM	Normal	100	.95	**.94**	.94	.95	**.87**	**.91**	.94	**.92**	**.93**	.95
BM	Normal	250	.95	**.93**	.94	.95	**.77**	**.87**	**.94**	**.91**	**.93**	.95
BM	Normal	1,000	.95	**.91**	**.94**	.95	**.37**	**.68**	**.92**	**.81**	**.90**	.94
BM	Normal	10,000	.95	**.60**	**.89**	.95	**.00**	**.01**	**.67**	**.12**	**.57**	**.92**
BM	Normal mixture	50	.95	.95	.95	.94	**.93**	**.93**	.95	**.93**	.95	.94
BM	Normal mixture	100	.95	.95	.94	**.93**	**.93**	**.94**	.95	**.93**	.94	**.93**
BM	Normal mixture	250	.95	.95	**.93**	**.89**	**.92**	**.93**	.94	**.92**	**.93**	**.89**
BM	Normal mixture	1,000	.95	.95	**.84**	**.68**	**.91**	**.94**	**.92**	**.84**	**.90**	**.70**
BM	Normal mixture	10,000	.95	**.92**	**.14**	**.00**	**.79**	**.94**	**.66**	**.19**	**.46**	**.00**
BM	Log-normal	50	.95	**.86**	**.89**	**.92**	**.67**	**.73**	**.83**	**.90**	**.92**	**.94**
BM	Log-normal	100	.95	**.79**	**.84**	**.90**	**.43**	**.53**	**.72**	**.86**	**.90**	**.93**
BM	Log-normal	250	.95	**.57**	**.70**	**.84**	**.08**	**.17**	**.43**	**.75**	**.85**	**.92**
BM	Log-normal	1,000	.95	**.07**	**.21**	**.55**	**.00**	**.00**	**.01**	**.33**	**.60**	**.84**
BM	Log-normal	10,000	.95	**.00**	**.00**	**.00**	**.00**	**.00**	**.00**	**.00**	**.00**	**.18**

Note. Coverage probabilities are boldface if they are substantially too low (< 0.94). N = sample size; BM = Brunner–Munzel's test. None, IQR, MAD, and *z*-score were the outlier removal strategies. IQR = interquartile range; MAD = median absolute deviation. The numbers (1.5, 2, 3) refer to the cutoff used by the respective outlier removal approach.

(*Appendices continue*)

Table A5

Relative Biases When Removing Outliers Across Groups; Unequal Variances

				IQR			MAD			*z*-score		
Test	Data	*N*	None	1.5	2	3	1.5	2	3	1.5	2	3
Welch	Normal	50	0.00	**−0.16**	**−0.08**	**−0.01**	**−0.52**	**−0.35**	**−0.14**	**−0.35**	**−0.20**	**−0.05**
Welch	Normal	100	0.00	**−0.16**	**−0.07**	−0.01	**−0.51**	**−0.35**	**−0.13**	**−0.35**	**−0.20**	**−0.05**
Welch	Normal	250	0.00	**−0.15**	**−0.07**	−0.01	**−0.51**	**−0.34**	**−0.13**	**−0.35**	**−0.20**	**−0.05**
Welch	Normal	1,000	0.00	**−0.15**	**−0.07**	−0.01	**−0.51**	**−0.34**	**−0.13**	**−0.35**	**−0.20**	**−0.05**
Welch	Normal	10,000	0.00	**−0.15**	**−0.06**	−0.01	**−0.51**	**−0.34**	**−0.13**	**−0.35**	**−0.20**	**−0.05**
Welch	Normal mixture	50	0.00	**−0.46**	**−0.53**	**−0.57**	**−0.59**	**−0.50**	**−0.49**	**−0.37**	**−0.49**	**−0.54**
Welch	Normal mixture	100	0.00	**−0.45**	**−0.53**	**−0.59**	**−0.59**	**−0.49**	**−0.48**	**−0.35**	**−0.49**	**−0.57**
Welch	Normal mixture	250	0.00	**−0.44**	**−0.53**	**−0.59**	**−0.59**	**−0.49**	**−0.48**	**−0.34**	**−0.48**	**−0.58**
Welch	Normal mixture	1,000	0.00	**−0.44**	**−0.52**	**−0.60**	**−0.59**	**−0.49**	**−0.47**	**−0.33**	**−0.48**	**−0.59**
Welch	Normal mixture	10,000	0.00	**−0.43**	**−0.52**	**−0.60**	**−0.59**	**−0.49**	**−0.47**	**−0.33**	**−0.48**	**−0.59**
Welch	Log-normal	50	0.00	**−1.06**	**−0.96**	**−0.79**	**−1.31**	**−1.27**	**−1.14**	**−0.93**	**−0.80**	**−0.59**
Welch	Log-normal	100	0.00	**−1.06**	**−0.96**	**−0.79**	**−1.32**	**−1.27**	**−1.14**	**−0.92**	**−0.79**	**−0.58**
Welch	Log-normal	250	0.00	**−1.06**	**−0.96**	**−0.79**	**−1.33**	**−1.27**	**−1.14**	**−0.91**	**−0.77**	**−0.57**
Welch	Log-normal	1,000	0.00	**−1.06**	**−0.96**	**−0.79**	**−1.33**	**−1.28**	**−1.14**	**−0.90**	**−0.76**	**−0.55**
Welch	Log-normal	10,000	0.00	**−1.06**	**−0.96**	**−0.79**	**−1.34**	**−1.28**	**−1.15**	**−0.89**	**−0.76**	**−0.55**
BM	Normal	50	0.00	**−0.01**	−0.00	−0.00	**−0.04**	**−0.03**	−0.01	**−0.02**	**−0.01**	−0.00
BM	Normal	100	0.00	−0.01	−0.00	−0.00	**−0.04**	**−0.03**	−0.01	**−0.02**	−0.01	−0.00
BM	Normal	250	0.00	−0.01	−0.00	−0.00	**−0.04**	**−0.03**	−0.01	**−0.02**	−0.01	−0.00
BM	Normal	1,000	0.00	−0.01	−0.00	−0.00	**−0.04**	**−0.03**	−0.01	**−0.02**	−0.01	−0.00
BM	Normal	10,000	0.00	−0.01	−0.00	−0.00	**−0.04**	**−0.03**	−0.01	**−0.02**	−0.01	−0.00
BM	Normal mixture	50	0.00	−0.01	**−0.02**	**−0.02**	−0.01	−0.00	**−0.01**	**0.01**	**−0.01**	**−0.02**
BM	Normal mixture	100	0.00	−0.00	**−0.02**	**−0.02**	−0.01	−0.00	−0.01	**0.01**	**−0.01**	**−0.02**
BM	Normal mixture	250	0.00	−0.00	**−0.02**	**−0.02**	−0.01	−0.00	−0.01	**0.01**	**−0.01**	**−0.02**
BM	Normal mixture	1,000	0.00	−0.00	**−0.02**	**−0.02**	−0.01	−0.00	−0.01	**0.02**	**−0.01**	**−0.02**
BM	Normal mixture	10,000	0.00	−0.00	**−0.02**	**−0.02**	−0.01	−0.00	−0.01	**0.02**	**−0.01**	**−0.02**
BM	Log-normal	50	0.00	**−0.06**	**−0.05**	**−0.03**	**−0.13**	**−0.11**	**−0.08**	**−0.05**	**−0.03**	**−0.02**
BM	Log-normal	100	0.00	**−0.06**	**−0.05**	**−0.03**	**−0.13**	**−0.11**	**−0.08**	**−0.05**	**−0.03**	**−0.02**
BM	Log-normal	250	0.00	**−0.06**	**−0.05**	**−0.03**	**−0.13**	**−0.11**	**−0.08**	**−0.04**	**−0.03**	**−0.02**
BM	Log-normal	1,000	0.00	**−0.06**	**−0.05**	**−0.03**	**−0.13**	**−0.11**	**−0.08**	**−0.04**	**−0.03**	**−0.02**
BM	Log-normal	10,000	0.00	**−0.06**	**−0.05**	**−0.03**	**−0.13**	**−0.11**	**−0.08**	**−0.04**	**−0.03**	**−0.02**

Note. Relative bias estimates are boldface if they are substantially different from zero (< 0.01 or > 0.01). $N =$ sample size; BM = Brunner–Munzel's test. None, IQR, MAD, and *z*-score were the outlier removal strategies. IQR = interquartile range; MAD = median absolute deviation. The numbers (1.5, 2, 3) refer to the cutoff used by the respective outlier removal approach.

(*Appendices continue*)

Table A6
Power When Removing Outliers Across Groups; Unequal Variances

				IQR			MAD			*z*-score		
Test	Data	*N*	None	1.5	2	3	1.5	2	3	1.5	2	3
Welch	Normal	50	0.70	**.62**	**.67**	.70	**.38**	**.50**	**.63**	**.54**	**.62**	**.68**
Welch	Normal	100	.94	**.89**	**.92**	.94	**.63**	**.78**	**.90**	**.82**	**.88**	**.93**
Welch	Normal	250	1.00	1.00	1.00	1.00	**.94**	**.99**	1.00	**.99**	1.00	1.00
Welch	Normal	1,000	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Welch	Normal	10,000	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Welch	Normal mixture	50	.38	**.29**	**.21**	**.16**	**.28**	**.32**	**.25**	.40	**.25**	**.17**
Welch	Normal mixture	100	.69	**.52**	**.36**	**.26**	**.47**	**.55**	**.45**	.69	**.43**	**.27**
Welch	Normal mixture	250	.98	**.90**	**.71**	**.52**	**.83**	**.90**	**.84**	**.97**	**.78**	**.54**
Welch	Normal mixture	1,000	1.00	1.00	1.00	**.98**	1.00	1.00	1.00	1.00	1.00	**.98**
Welch	Normal mixture	10,000	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Welch	Log-normal	50	.18	**.02**	**.03**	**.06**	**.00**	**.00**	**.01**	**.04**	**.06**	**.11**
Welch	Log-normal	100	.39	**.02**	**.04**	**.10**	**.00**	**.00**	**.01**	**.05**	**.10**	**.19**
Welch	Log-normal	250	.79	**.01**	**.05**	**.18**	**.00**	**.00**	**.00**	**.09**	**.20**	**.44**
Welch	Log-normal	1,000	1.00	**.01**	**.08**	**.55**	**.00**	**.00**	**.00**	**.22**	**.62**	**.95**
Welch	Log-normal	10,000	1.00	**.00**	**.30**	1.00	**.00**	**.00**	**.00**	**.93**	1.00	1.00
BM	Normal	50	.65	**.58**	**.62**	.65	**.36**	**.47**	**.59**	**.49**	**.57**	**.63**
BM	Normal	100	.92	**.86**	**.89**	**.91**	**.60**	**.75**	**.87**	**.77**	**.85**	**.90**
BM	Normal	250	1.00	1.00	1.00	1.00	**.93**	**.98**	1.00	**.99**	1.00	1.00
BM	Normal	1,000	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
BM	Normal	10,000	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
BM	Normal mixture	50	.33	**.29**	**.24**	**.22**	**.26**	**.29**	**.26**	.36	**.27**	**.22**
BM	Normal mixture	100	.57	**.51**	**.43**	**.38**	**.43**	**.50**	**.47**	.62	**.47**	**.39**
BM	Normal mixture	250	.92	**.88**	**.81**	**.74**	**.78**	**.86**	**.85**	.95	**.84**	**.74**
BM	Normal mixture	1,000	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
BM	Normal mixture	10,000	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
BM	Log-normal	50	.03	.15	.11	.08	.35	.29	.18	.11	.08	.05
BM	Log-normal	100	.03	.24	.17	.11	.60	.50	.31	.16	.11	.07
BM	Log-normal	250	.04	.48	.34	.19	.94	.86	.63	.29	.18	.10
BM	Log-normal	1,000	.05	.96	.86	.56	1.00	1.00	.99	.76	.52	.24
BM	Log-normal	10,000	.15	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	.96

Note. Power estimates are boldface if removing outliers decreased the power. *N* = sample size; BM = Brunner–Munzel's test. None, IQR, MAD, and *z*-score were the outlier removal strategies. IQR = interquartile range; MAD = median absolute deviation. The numbers (1.5, 2, 3) refer to the cutoff used by the respective outlier removal approach.

(*Appendices continue*)

Appendix B

Results of Comparing Groups: Null Hypothesis True

Table B1

Type I Error Rates When Removing Outliers Across Groups

				IQR			MAD			*z*-score		
Test	Data	*N*	None	1.5	2	3	1.5	2	3	1.5	2	3
Welch	Normal	50	.05	**.06**	.06	.05	**.07**	**.07**	.06	**.07**	**.07**	.06
Welch	Normal	100	.05	**.07**	.06	.05	**.07**	**.07**	.06	**.07**	**.07**	.06
Welch	Normal	250	.05	**.07**	.06	.05	**.07**	**.07**	.06	**.07**	**.07**	.06
Welch	Normal	1,000	.05	**.07**	.06	.05	**.07**	**.07**	.06	**.08**	**.07**	.06
Welch	Normal	10,000	.05	**.07**	.06	.05	**.07**	**.07**	.06	**.08**	**.07**	.06
Welch	Normal mixture	50	.05	**.12**	**.17**	**.23**	**.09**	**.10**	**.14**	**.09**	**.16**	**.21**
Welch	Normal mixture	100	.05	**.19**	**.30**	**.42**	**.10**	**.13**	**.23**	**.11**	**.26**	**.39**
Welch	Normal mixture	250	.05	**.37**	**.61**	**.82**	**.15**	**.22**	**.47**	**.16**	**.51**	**.79**
Welch	Normal mixture	1,000	.05	**.88**	**.99**	**1.00**	**.35**	**.59**	**.96**	**.38**	**.97**	**1.00**
Welch	Normal mixture	10,000	.05	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
Welch	Log-normal	50	**.07**	**.75**	**.63**	**.42**	**.84**	**.93**	**.78**	**.58**	**.43**	**.25**
Welch	Log-normal	100	**.06**	**.95**	**.87**	**.66**	**.98**	**1.00**	**.97**	**.80**	**.62**	**.36**
Welch	Log-normal	250	.06	**1.00**	**1.00**	**.95**	**1.00**	**1.00**	**1.00**	**.98**	**.90**	**.61**
Welch	Log-normal	1,000	.05	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**.97**
Welch	Log-normal	10,000	.05	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
BM	Normal	50	.05	**.06**	.05	.05	**.07**	**.07**	.06	**.07**	**.07**	.05
BM	Normal	100	.05	**.06**	.05	.05	**.07**	**.07**	.06	**.07**	**.07**	.05
BM	Normal	250	.05	**.06**	.05	.05	**.07**	**.07**	.06	**.07**	**.06**	.05
BM	Normal	1,000	.05	**.06**	.05	.05	**.07**	**.07**	.06	**.07**	**.07**	.05
BM	Normal	10,000	.05	**.06**	.05	.05	**.07**	**.07**	.06	**.07**	**.07**	.05
BM	Normal mixture	50	.05	**.06**	.06	**.07**	**.11**	**.08**	.06	**.07**	.06	**.06**
BM	Normal mixture	100	.05	**.06**	**.06**	**.08**	**.14**	**.09**	.06	**.08**	**.06**	**.08**
BM	Normal mixture	250	.05	**.06**	**.07**	**.14**	**.24**	**.11**	.06	**.12**	**.07**	**.13**
BM	Normal mixture	1,000	.05	**.06**	**.13**	**.40**	**.64**	**.25**	**.07**	**.28**	**.08**	**.39**
BM	Normal mixture	10,000	.05	**.09**	**.72**	**1.00**	**1.00**	**.96**	**.15**	**.99**	**.33**	**1.00**
BM	Log-normal	50	.05	**.09**	**.08**	**.06**	**.22**	**.13**	**.09**	**.08**	**.07**	.06
BM	Log-normal	100	.05	**.14**	**.11**	**.08**	**.25**	**.20**	**.13**	**.11**	**.08**	**.06**
BM	Log-normal	250	.05	**.27**	**.21**	**.13**	**.28**	**.42**	**.28**	**.18**	**.12**	**.07**
BM	Log-normal	1,000	.05	**.78**	**.62**	**.36**	**.34**	**.95**	**.80**	**.50**	**.31**	**.13**
BM	Log-normal	10,000	.05	**1.00**	**1.00**	**1.00**	**.72**	**1.00**	**1.00**	**1.00**	**.99**	**.71**

Note. Type I error rates are boldface if they are substantially too large (> .06). *N* = sample size; BM = Brunner–Munzel's test. None, IQR, MAD, and *z*-score were the outlier removal strategies. IQR = interquartile range; MAD = median absolute deviation. The numbers (1.5, 2, 3) refer to the cutoff used by the respective outlier removal approach.

(*Appendices continue*)

Table B2
Biases When Removing Outliers Across Groups; Null Hypothesis True

				IQR			MAD			*z*-score		
Test	Data	*N*	None	1.5	2	3	1.5	2	3	1.5	2	3
Welch	Normal	50	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Welch	Normal	100	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Welch	Normal	250	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Welch	Normal	1,000	0.00	0.00	−0.00	−0.00	0.00	0.00	0.00	−0.00	0.00	−0.00
Welch	Normal	10,000	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Welch	Normal mixture	50	0.00	**0.22**	**0.33**	**0.44**	**0.08**	**0.14**	**0.27**	**0.13**	**0.30**	**0.41**
Welch	Normal mixture	100	0.00	**0.22**	**0.33**	**0.45**	**0.08**	**0.14**	**0.27**	**0.12**	**0.29**	**0.43**
Welch	Normal mixture	250	0.00	**0.22**	**0.33**	**0.46**	**0.08**	**0.13**	**0.27**	**0.11**	**0.29**	**0.45**
Welch	Normal mixture	1,000	0.00	**0.21**	**0.33**	**0.46**	**0.08**	**0.13**	**0.26**	**0.11**	**0.29**	**0.45**
Welch	Normal mixture	10,000	0.00	**0.21**	**0.33**	**0.46**	**0.08**	**0.13**	**0.26**	**0.10**	**0.29**	**0.46**
Welch	Log-normal	50	0.00	**0.76**	**0.70**	**0.59**	**0.71**	**0.83**	**0.77**	**0.67**	**0.59**	**0.44**
Welch	Log-normal	100	0.00	**0.76**	**0.70**	**0.59**	**0.72**	**0.84**	**0.77**	**0.67**	**0.58**	**0.43**
Welch	Log-normal	250	0.00	**0.76**	**0.70**	**0.59**	**0.73**	**0.85**	**0.77**	**0.66**	**0.57**	**0.42**
Welch	Log-normal	1,000	0.00	**0.76**	**0.70**	**0.60**	**0.74**	**0.85**	**0.77**	**0.65**	**0.56**	**0.42**
Welch	Log-normal	10,000	0.00	**0.76**	**0.70**	**0.60**	**0.74**	**0.85**	**0.77**	**0.65**	**0.56**	**0.41**
BM	Normal	50	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
BM	Normal	100	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
BM	Normal	250	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
BM	Normal	1,000	0.00	0.00	0.00	−0.00	0.00	0.00	0.00	0.00	0.00	0.00
BM	Normal	10,000	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
BM	Normal mixture	50	0.00	−0.00	**0.01**	**0.02**	**−0.04**	**−0.02**	0.00	**−0.02**	0.01	**0.02**
BM	Normal mixture	100	0.00	−0.00	**0.01**	**0.02**	**−0.04**	**−0.02**	0.00	**−0.02**	0.01	**0.02**
BM	Normal mixture	250	0.00	−0.00	**0.01**	**0.02**	**−0.04**	**−0.02**	0.00	**−0.02**	0.01	**0.02**
BM	Normal mixture	1,000	0.00	−0.00	**0.01**	**0.02**	**−0.04**	**−0.02**	0.00	**−0.02**	0.01	**0.02**
BM	Normal mixture	10,000	0.00	−0.00	**0.01**	**0.02**	**−0.04**	**−0.02**	0.00	**−0.02**	0.01	**0.02**
BM	Log-normal	50	0.00	**0.04**	**0.03**	**0.02**	−0.00	**0.05**	**0.04**	**0.03**	**0.02**	**0.01**
BM	Log-normal	100	0.00	**0.04**	**0.03**	**0.02**	0.00	**0.05**	**0.04**	**0.03**	**0.02**	**0.01**
BM	Log-normal	250	0.00	**0.04**	**0.03**	**0.02**	0.01	**0.05**	**0.04**	**0.03**	**0.02**	**0.01**
BM	Log-normal	1,000	0.00	**0.04**	**0.03**	**0.02**	**0.01**	**0.05**	**0.04**	**0.03**	**0.02**	**0.01**
BM	Log-normal	10,000	0.00	**0.04**	**0.03**	**0.02**	**0.01**	**0.05**	**0.04**	**0.03**	**0.02**	**0.01**

Note. Bias estimates are boldface if they are substantially different from zero (< 0.01 or > 0.01). N = sample size; BM = Brunner–Munzel's test. None, IQR, MAD, and *z*-score were the outlier removal strategies. IQR = interquartile range; MAD = median absolute deviation. The numbers (1.5, 2, 3) refer to the cutoff used by the respective outlier removal approach.

(*Appendices continue*)

Appendix C

Results of Regression

Table C1

Coverage Probabilities for Hypothesis-Blind Outlier Removal in a Regression Model

Data DV	Data IV	*N*	None	Blind
Normal	Normal	50	.95	**.84**
Normal	Normal	100	.95	**.75**
Normal	Normal	250	.95	**.50**
Normal	Normal	1,000	.95	**.03**
Normal	Normal	10,000	.95	**.00**
Normal	Categorical	50	.95	**.86**
Normal	Categorical	100	.95	**.82**
Normal	Categorical	250	.95	**.71**
Normal	Categorical	1,000	.95	**.29**
Normal	Categorical	10,000	.95	**.00**
Normal mixture	Normal	50	.95	**.89**
Normal mixture	Normal	100	.95	**.85**
Normal mixture	Normal	250	.95	**.72**
Normal mixture	Normal	1,000	.95	**.24**
Normal mixture	Normal	10,000	.95	**.00**
Normal mixture	Categorical	50	.95	**.92**
Normal mixture	Categorical	100	.95	**.90**
Normal mixture	Categorical	250	.95	**.84**
Normal mixture	Categorical	1,000	.95	**.56**
Normal mixture	Categorical	10,000	.95	**.00**
Log-normal	Normal	50	.95	**.92**
Log-normal	Normal	100	.95	**.90**
Log-normal	Normal	250	.95	**.85**
Log-normal	Normal	1,000	.95	**.64**
Log-normal	Normal	10,000	.95	**.00**
Log-normal	Categorical	50	.96	**.93**
Log-normal	Categorical	100	.96	**.91**
Log-normal	Categorical	250	.95	**.87**
Log-normal	Categorical	1,000	.95	**.66**
Log-normal	Categorical	10,000	.95	**.00**

Note. Coverage probabilities are boldface if they are substantially too low (< 0.94). N = sample size; DV = dependent variable; IV = independent variable.

Table C2

Relative Biases for Hypothesis-Blind Outlier Removal in a Regression Model

Data DV	Data IV	*N*	None	Blind
Normal	Normal	50	.00	**−.13**
Normal	Normal	100	.00	**−.13**
Normal	Normal	250	.00	**−.13**
Normal	Normal	1,000	.00	**−.13**
Normal	Normal	10,000	.00	**−.13**
Normal	Categorical	50	.00	**−.08**
Normal	Categorical	100	.00	**−.08**
Normal	Categorical	250	.00	**−.08**
Normal	Categorical	1,000	.00	**−.08**
Normal	Categorical	10,000	.00	**−.08**
Normal mixture	Normal	50	.00	**−.11**
Normal mixture	Normal	100	.00	**−.10**
Normal mixture	Normal	250	.00	**−.10**
Normal mixture	Normal	1,000	.00	**−.09**
Normal mixture	Normal	10,000	.00	**−.09**
Normal mixture	Categorical	50	.00	**−.08**
Normal mixture	Categorical	100	.00	**−.07**
Normal mixture	Categorical	250	.00	**−.07**
Normal mixture	Categorical	1,000	.00	**−.06**
Normal mixture	Categorical	10,000	.00	**−.06**
Log-normal	Normal	50	.00	**−.08**
Log-normal	Normal	100	.00	**−.07**
Log-normal	Normal	250	.00	**−.07**
Log-normal	Normal	1,000	.00	**−.06**
Log-normal	Normal	10,000	.00	**−.06**
Log-normal	Categorical	50	.00	**−.07**
Log-normal	Categorical	100	.00	**−.07**
Log-normal	Categorical	250	.00	**−.06**
Log-normal	Categorical	1,000	.00	**−.06**
Log-normal	Categorical	10,000	.00	**−.06**

Note. Relative bias estimates are boldface if they are substantially different from zero (< -0.01 or > 0.01). N = sample size; DV = dependent variable; IV = independent variable.

Received January 21, 2022
Revision received November 18, 2022
Accepted December 01, 2022 ■