**Reviewer 1 :**

**Comment 1) l. 62-66   
“Although not the focus of the current article, additional tests exist that allow researchers to examine hypotheses about other relevant parameters of distribution than the mean (such as standard deviations and the shape of the distribution (see for example Erceg-Hurn & Mirosevich, 2008; Wilcox, 1998)” In this text there are two left and one right parentheses, so a correction is needed.**

* **W*e removed the first left parenthese.***

**Comment 2) More importantly, I’m not sure I understand what the authors mean. The references are contributions to robust statistics, but the text does not seem in line with the references. Perhaps the authors mean that a robust statistics approach would use different estimators, such as the median or trimmed/winsorized mean, of the same parameter as in “standard” statistics (a measure of central tendency) rather than different parameters. Or perhaps they mean that by inspecting and testing other parameters (shape, variance) of the distribution, one could decide to resort to a non-parametric/robust statistics approach (as they argue at the end of the manuscript). The current sentence, however, is unclear on what they mean. Just to be clear, I’m not questioning the decision to focus on parametric statistics, which is legitimate. I’m simply asking to revise and improve the sentence.**

* ***You are perfectly right, the references are not in line with the text. We meant that comparing groups based on their means is not the only solution, other tests are based on different null hypothesis. Nonparametric tests compare the whole distributions (mean, sd, shape, etc.) and robust statistics use different measure of central tendency than the mean (median or trimmed : winsorized mean). We hope that the new sentence is clearer :***

***« Although not the focus of the current article, additional tests exist that allow researchers to compare groups either based on other estimators of central tendency than the mean (see for example Erceg-Hurn & Mirosevich, 2008; Wilcox, 1998), or based on other relevant parameters of distribution than the central tendency, such as standard deviations and the shape of the distribution (Tomarken & Serlin, 1986) »***

**Comment 3) l. 165 “Consequences of Assumption Violations” should be in bold (there must have been a problem in formatting the text).**

* ***It was indeed a problem in formatting the text and it is now corrected.***

**l. 307 Excel**

* ***We corrected this typo, as well as another one in the first endnote : « Note that this is a didactic example ».***

**Reviewer 2**

**Comment 1) On p.13, in the simulation description, and for consistency, the notation for SD-ratio and n-ratio should be introduced in the same part of the paragraph.**

* ***We introducted both notations in the same part of the paragraph : « Based on the literature review presented above, we manipulated the number of groups, the sample sizes, the sample size ratio (n-ratio = ), the SD-ratio (SD-ratio =), and the sample size and variance pairing ».***

**Comment 2) Is the definition of positive/negative pairing on p.13 correct? It seems counterintuitive, covers twice the same case (positively correlated SD and n), and incongruent with the**

**use of correlations later in the paper.**

* ***The definition of negative pairing is indeed wrong. By negative pairing, we mean that the group with the ~~smallest~~ largest sample size is extracted from the population with the smallest SD.***

**Comment 3) On figures 2-6, ‘alpha’ should be replaced by Type I error rate, as you explicitly define alpha as the nominal error rate, while you observed the empirical one through simulation, and are interested in departures from the nominal level. This might confuse some readers.**

* ***We have changed the script in order to replace « alpha » by Type I error rate. We agree that it might confuse some readers.***

**Comment 4) On the same figures, the vertical scale as been kept constant across all figures for comparison, simply due to figure 6 for the F-test case. Even so, the values are out of scale and hard to read, so the figures could overall be improved for readability (although I’m usually agnostic to the package used for graphics, I think here relying on ‘ggplot2’ instead of base ‘graphics’ would help).**

* ***In order to avoid that values are out of scale, we changed the vertical scale (going now from .02 to .14 in all figures). Concerning the readability of the graph, our goal is to show a general tendency of the Type I error rate and to show in which situations the type I error rate is within (or outside) the acceptable range of type I error rate, considering Bradley’s criterions (1978). The exact values of the means are quite anecdotic to convey this message.***

**Comment 5) Graph legends (SDs) are partially missing on all Figure A1-6 (which density correspond to which parameter value). On Figure A1-6, there is no “function of the population SD”, and this must be an approximated/estimated density, since the chisquared reaches negative values (e.g., through plot(density(rchisq(…))) instead of plot(dchisq(…) in R).**

* ***Graph legends are now visible. Indeed, we used plot(density(rdistribution(…))) instead of plot(ddistribution), which was time consuming. The code in the .rmd file of the appendix was modified accordingly.***

**Comment 6) In the final recommendations (p.31), the sentence “Unless you have good reasons to believe that distributions underlying the data have small kurtosis and skewness, we recommend to avoid alternative tests that are not based on means comparison” is misleading. Such tests may include non-parametric tests (indicated afterwards) and parametric tests specially designed to handle such distributions and the limits of W-test (e.g., for asymmetric or kurtoleptic distributions, as they are widespread in real datasets).**

* ***We understand the reviewer’s confusion, as there is a typo in our original sentence. Indeed, the right sentence is : « we recommend to avoid alternative tests that are ~~not~~ based on means comparison ». This is corrected in the new file.***

**Comment 7) Again in recommendations, the last sentence “For more information, see Erceg-Hurn and Mirosevich (2008)” indicates that they is more to read in the reference than about trimmed means tests, but the article does not indicate what other elements from the recommendations or issues at hand are addressed (the reader could guess “when the focus is on other parameters than the mean” as the reference is provided in this context in other parts of the paper, but this remains doubtful given the previous sentence)**

* ***We modified the sentence as following : « For more information about robust alternatives that are based on other parameters than the mean, see Erceg-Hurn & Mirosevich (2008) ».***

**Comment 8) When I wrote the review, I had “response time” in mind for RT, as a more general term encompassing “reaction time” on specific tasks. Thus, I don’t see why it should be bound to “neurosciences” (especially with the psychology audience targeted). Apart from that, thank you for considering my remark and even including the example in the paper.**

* ***We replaced « Reaction time » with « Response time » as following : « ~~Such examples can be found in the ﬁeld of neurosciences such as reaction times, that~~ For example, response times can be very large but never below zero (resulting in right-skewed distributions ». We also removed the reference of Ratcliff which was bound to the context of reaction times.***