Dear Prof. Dompnier,

Thank you very much for this review that we deemed of high quality. Please find our responses below, we hope that this improvement reaches the standards of qualilty of the IRSP. We tried to thouroughly rework the paper in the aim of being much more didactic for psychologists.

Editor’s comment:

Whereas Reviewer 1 thinks your paper is potentially publishable pending major revision, Reviewer 2 has some doubts about its suitability for publication. Based on my own reading, I share most of the Reviewers’ concerns but think that your paper could be an interesting contribution after an extensive revision. Accordingly, I am inviting you to revise your paper and resubmit it for further consideration.

**Comment 1) To sum up, Reviewer 1 thinks that some conclusions are not fully supported by the data and that some (important) results are not discussed in the document. I would add that I personally found difficult to identify the relevant results in the Tables. I thus think that the presentation of results and their discussion should be deeply modified.**

***We fully reworked the discussion and the text accordingly, added graphs and tables in order to improve its readability.***

**Comment 2) Reviewer 2 points out that the document contains misconceptions or conceptual errors that undermine the pedagogical aim of the paper. In this regard, I would add that the International Review of Social Psychology publishes methodological articles that present in a pedagogical way statistical tools useful to social psychologists. In my opinion, the manuscript is currently quite difficult for non-specialists to read (especially the results). In other words, I think that the manuscript should be substantially modified in order to increase its pedagogical usefulness**.

***We agree with this comment and took it into account as mentioned above.***

Reviewer 1:  
  
I have read this manuscript with interest. It is a technically competent contribution providing practically relevant suggestions. Despite many positive features, I think that there are few major issues that need to be addressed. The most important of them all is the key message of this contribution: after reading this paper, as well as other relevant papers on the same or related issues, I am not convinced that one should always use the W-test instead of the F-test.

The authors provide a compelling introduction and present very well their contribution. They should also be complimented in the transparency of the contribution and in making available the more extensive set of results in the SM and as additional material in the OSF.

**Comment 1) In my opinion, the main problem is in the discussion of (some) of the results. After having inspected the Excel table of the Type I error in the OSF as referred to in the SM, I am not sure that the main gist (the three main points) in the manuscript is an accurate reflection of some of the results. My reading of the Excel table is that when the homoscedasticity assumption is respected (i.e., variances are equal across groups), the F-test always outperforms the W-test (and the F\* test).**

***We carefully revised our interpretation of the data. We believe that (unsurprisingly) when all assumption are met, the F-test slightly outperforms the W-test (and the F\*-test). We didn’t underline this issue in the first version, but we address this more carefully this time by stating:***

* ***About the Type I error: “The F-test and F\*-test only marginally deviates from the nominal 5%, regardless the underlying distribution and the SD-ratio. The W-tests also only marginally deviates from the nominal 5%, except under asymmetry (the tests becomes a little more sensitive) or extremely heavy tails (the test becomes more conservative), consistently with Harwell et al. (1992). However, deviations don’t exceed the liberal criterion of Bradley (1978). »* (lines 348 – 353)**
* ***About the power: “When distributions are normal, the W-test is slightly less powerful than the F-test and F\*-test, even if differences are very small (always smaller than 3%).” (*lines 402 – 404)**

**Comment 2) The gap in performance increases as the number of groups grow and holds regardless of whether the sample sizes are unequal (i.e., for both balanced and unbalanced designs) and for all types of distribution tested.**

***After consideration we believe that when the number of groups grow, the tests become more liberal regardless of the condition. Since there is no interaction effect between the number of groups and the condition, we just underlined the main effect of the group and only report the three groups condition.***

***“For didactic reasons, we will report only the results when we compare three groups (k=3). Increasing the number of groups have as main effect to increase the liberality of all test. For interested readers, […]* (lines 305 – 309).**

**Comment 3) The gap tends to increase with higher skewness, both positive and negative, with more groups, and with smaller sample sizes. These results are not reported nor discussed in the main manuscript.**

***In the new graphs, we systematically showed the difference between the W-test, F-test and F\*-test for each simulated distribution underlying the data. This allows to more clearly illustrate whether the consequences of different conditions (heteroscedasticity, balanced vs. unbalanced designs, etc.) are similar under all distributions.***

***Regarding the sample size effect we collapsed all sample sizes in the graphs, for the sake of clarity. However, we discuss the sample size effect in the following sentence: “Overall, the higher the sample sizes, the less the distributions of the population underlying the samples impact the robustness of the tests [@Srivastava\_1959]. However, increasing the sample sizes does not improve the robustness of the test when there is heteroscedasticity. Interested reader can see all details in the following Excell spreadsheet, available on github : « Type I error rate.xlsx ».”* (lines 310 – 314).**

The second point presented is a correct description of the results, that is the Type I error of the F-test is heavily affected by heteroscedasticity, with both over and underestimation depending on the direction of the correlation between N and SD. The W-test is better in this respect. 

**Comment 4) Concerning the third point (for highly skewed distributions the W-test is better), I could not see in the Excel table results supporting it. The results reported in the manuscript (Table 4) is about one combination of highly skewed scores with unequal variances, which is not the same as the point made. I believe that these results may be driven by unequal variances (in line with the results summed up in point 2), regardless of whether the design is balanced. Perhaps I have missed something in the Excel table, or perhaps the results are not reported there, but I could not see the summary results of a simulation with high skewness, but equal variances, allowing to make this point.**

***We are now more exhaustive in the description of these results, having introduced 9 sub-conditions that we discussed based on new graphs (see Figure 1, 2 and 3 and its description)***

**Comment 5) The results for power are presented better in the corresponding table in the OSF than in the manuscript. As for the Type I error, the easier to understand presentation is in terms of reference with theoretical power rather than comparing the tests. This latter way of presenting the results does not allow to understand easily which test performs better. Moreover, it is inconsistent with how results are presented for Type I error. I think that the tables should be corrected and the discussion revised.**

***We finally choose to present the results for power both in terms of raw power and consistency (i.e. in terms of reference with theoretical power, as you suggested).***

**Comment 6) Anyway, the discussion concerning the results for power is slightly more balanced but still reflects only partly the full results in the Excel tables in OSF. The F test performs almost always better than the W test for three groups, even in the presence of heteroscedasticity. In some cases, the difference is substantial. The gap in performance tends to diminish or disappear with increasing sample sizes. These results are not mentioned in the main text. For two groups the results are more balanced.**

***This misinterpretation is due to ambiguity in the way we transmitted results. Because we wanted to see the impact of non-normality independently of the impact of heteroscedasticity, we did not compute (O-E)/E in folders dedicated to non-normal distributions. We rather compared observed power when distributions were normal or not (Odist-Onormal)/Onormal. Because we didn’t mention it, one could believe that the F-test is better than the W-test, under heteroscedasticity but it’s not the case: F-test is sometimes less affected by non-normal distributions than W-test, especially when there are heteroscedasticity, however, when both heteroscedasticity and non-normality are combined, W-test is way better than F-test.***

**Comment 7) The final recommendations, therefore, are a mixed bag. The first recommendation does not reflect the more nuanced results. In a nutshell, from my reading of the tables in OSF, the full results seem to suggest that the F-test performs better than the W test when variances are equal, both for Type I and power (not always for this latter, but almost invariably, and sometimes substantially, with three groups). The W test instead is generally better in the presence of heteroscedasticity, but not concerning power with three groups. I could not see clear results supporting the point that the W test performs better than the F test with high skewness when considered by itself and not in combination with other aspects.**

***We agree that the F-test is better than the W-test under the normality and homoscedasticity assumptions. However, differences are very small and both tests are very good when both assumptions are met. On the other side, when there is heteroscedasticity, the F-test is either too liberal or too conservative while the W-test is still good. In summary, W-test is almost always very good when the F-test performs well, and also very good when the F-test is flawed. As it is not possible to correctly detect departures from the normality and heteroscedasticity assumptions, W-test is a more viable solution than the F-test. Note that under specific configurations (e.g. small samples and highly skewed distributions combined with heteroscedasticity), both test perform badly. For these reason, we also gave a recommendation related to hypotheses that are not based on means.***

**Comment 8) These results suggest that there is no overall “winner”, but it depends on a combination of features. This is in contrast with the current main message of the contribution.**

***As we just mentioned, we believe that this is a question of opinion in the sense that indeed, the F-test outperforms the W-test in some configurations, however, we argue that in real situation it is difficult to accurately diagnose the given configuration. The W-test is only slightly less performant than the F-test, but always at least acceptable even when it is not the best solution. On the contrary, the F-test can become severely biased in some situations likely to be encounter in real settings. For these reasons, we recommend the use of the W-test by default.***

**Comment 9) The results also seem roughly in line with previous simulations and contributions, some of which have not been considered in this contribution (Harwell et al., 1992; Wilcox et al., 1986).**

***We took these suggested contributions into account (and thank the reviewer for this helpful suggestion), as well as other references cited in these papers.***

**Comment 10) Finally, concerning the issue of the problems in adopting a two-step approach to decide which test is better depending on the homoscedasticity assumption (pp. 3-4), there is a recent contribution suggesting a procedure that seems to represent a potentially viable solution (Kim & Cribbie, 2018).**

***Kim & Cribbie seem to think that people should always use W-test instead of F-test (but as they argue that this recommendation is unpopular, they show that an equivalence-based procedure in order to detect heteroscedasticity returns better results than the classical two-step procedures). Moreover, when analysing table 2 and 3 of their paper, one sees that in all their simulations, W-test (without gatekeeper) is equal or better than the two-step procedure using the equivalence-based procedure.***

**Comment 11) Incidentally, there is a reference to Figures 1 and 2 that is either misplaced, and there are no such figures, or the figures are missing (or anyway, I could not find them).**

***This has been changed accordingly (we removed the old Figures and introduced new ones).***

**Comment 12) To summarize, I like the basic idea of this work, and I appreciate the transparency in making available all results. However, I think that the contribution should be revised, most importantly by providing a much more nuanced picture of the results and henceforth more nuanced recommendations that line up better with them.**

***We revised accordingly.***

Reviewer 2:

**In the context of experimental psychology, the article underlines the need to not take linear model assumptions for granted, and not to stick to the classical F-test for one-way ANOVA (with F\* and W alternative tests considered). Monte Carlo simulations are run to add (computational) empirical evidence to past research and mathematical derivations in terms of type I and II error rates.  
  
The Welch adaptations of classical t-test and ANOVA have been available for years (even in SPSS) and walkthrough provided in many textbooks and online guides to applied statistics in psychology. The same is true for the recommendation of relying on descriptives to characterize shape and choose test accordingly, using Shapiro-Wilk for detecting departures from normality, or generally adopting a Neyman-Pearson approach to data testing (power analysis and thus control of type I/II error rates).  
  
Comment 1) Yet, the paper could be yet another useful attempt at pushing psychology researchers to switch to more flexible/robust methods, if it provided:  
- better articulation with the Delacre (2017) paper (since simply extending from the t-test to ANOVA)  
- comparison to nonparametric tests (as done in previous articles on the topic, including those from the same authors)  
- better referencing of past statistical research (of course from statistics, but also from empirical research methods themselves, even in psychology)**

***The text has been reworked in order to present the current hurdles, and stakes of the assumptions violations and we added several relevant refences (in line with reviewer 2 comment and reviewer 1 suggestions). However, we did not add simulations about nonparametric tests, nor other suggested solutions, for reasons explained below. We also added a paragraph in the paper in order to explain that we exclusively focus on hypotheses of mean comparisons: “Although not the focus of the current article, additional tests exist that allow researchers to examine hypotheses about other relevant parameters of a distribution than the mean (such as standard deviations and the shape of the distribution (see for example Erceg-Hurn & Mirosevich, 2008; Wilcox, 1998). However, since most researchers currenty generate hypotheses about differences between means [Erceg-Hurn & Mirosevich, 2008; Keselman et al., 1998], we think that a first realistic first step towards progress would be to get researchers to correctly test the hypothesis they are used to.” (lines 61 – 68).***

**Comment 2) Also, in its current state, there are many misconceptions (although often taken from more or less recent articles) as well as conceptual errors in the article, that may not serve its potential educational purpose. There are also missing figures, and the github source code does not serve the paper and replicability to its best.**

***We corrected the misconceptions (see below). Moreover, we tried to improve the github source in many ways:***

* ***Scripts were revised in order to be more efficient***
* ***A readme file was edited (to explain how we proceeded step by step)***
* ***Archives of previous submissions are available***
* ***In order that the reviewers can see how we took their comments into consideration, the draft version that as sent to the IRSP was moved into a Rmarkdown file. Each modification was then tracked on git.***

**Comment 3) Finally, in 2018, and even though t-test and ANOVA are still recurrent in experimental psychology, research methods papers should consider more general classes of models or methods. This is not only to consider more realistic designs in social psychology, but also for avoiding many of the pitfalls mentioned in the paper (especially since many of them are already used in social psychology):  
- linear model (minimally)  
- mixed models (allowing to control for more factors, thus preventing excessive mixed-normal distributions)  
- generalized models (for link functions and transformations avoiding both anormality and heteroscedasticity)  
- weighted regression (to prevent heteroscedasticity due to outliers)  
- Eicker or White-estimator (to get heteroscedasticity-consistent standard errors)**

***See our response in comment 1.***

I would therefore recommend rejecting the paper. To wrap up, for the paper to be really useful as yet another argument in favor of turning away from the old classical F-test for psychologists, it should get a bibliographical update, improved comparisons and simulation plan, as well as a user-friendly way of running simulation experiments and getting convinced by the results (not even thinking of R/Shiny interface or a code abstraction layer).

**Comment 4) "Parametric tests rely on two main assumptions: normality of the distribution and equality of variances." => This is plain wrong and should be reformuled. Within the very large class of parametric tests, this does not even hold for tests relying on the exponential family of probability distributions. And most importantly, it is of course not true for the Welch-adaptations, which are parametric.`**

***We changed accordingly (we do not refer to parametric tests in general but rather to t-test and ANOVA in particular).***   
  
**Comment 5) What is the data that should be normal for ANOVA? Not the IV of course, not the DV; only the DV conditioned by the IV(s), i.e. the residuals (as in any linear model)**

***Indeed, in the specific context of ANOVAs and t-test, the shape of the residuals distribution is similar to the shape of the data, because residuals are computed by substracting a constant to the DV scores. However, to avoid confusion we adjusted the text accordingly: “Student’s t-test and classical F-test ANOVA rely on the assumptions that two or more samples are independent and that independant and identically distributed residuals are normal and have equal variances between groups (in the abstract)”***

**Comment 6) Add i.i.d. to the clasical ANOVA assumptions (since independence is missing), even if not directly studied in this paper**

***We changed accordingly. In the abstract we mentioned all conditions, also in the first sentence of the introduction: “when comparing independent groups… rely on the assumptions that iid residuals… ”.***

**Comment 7) "quite robust conclusions" = in terms of type I or II errors ? This was already differentiated in Glass et al. (1972), and studied since then.**

***We systematically mentioned “robustness, in terms of Type I and either Type II error rates or power” (e.g., line 70: “does not strongly affect either the Type I error rates […] or* the power”).**

**Comment 8) "lack of attention to both the homoscedasticity and the normality assumptions" = how can this be inferred without error from non corrected dfs in papers? Since these corrections mostly are continuous and non linear, applying them by default may be the best option, but statistical softwares used in psychology and associated test procedures push to differentiate the reported statistics based on heteroscedasticity test results (e.g. Levene's based on mean, BrownñForsythe based on median). This sometimes leads to detectable inconsistencies between reported F, degrees of freedom and p-values, that may simply be due to applied correction (without reporting the adequate dfs)**

***We agree with your comment, however, we do not believe that it severely compromises the results of the investigation. Besides, this kind of inaccuracy also witnesses a lack of knowledge/attention of researches.***

**Comment 9) Why not considering nonparametric tests? They are also a standard alternative to violation of assumed distribution (here normality), often better deal with heteroscedasticity in terms of inference (e.g. old series of papers by Edgington and Onghena, even though distribution similarity across conditions remains an assumption), while being less powerful only in specific situations (small samples sizes which are unconventional anyway for parametric tests, or  perfect satisfaction of parametric assumptions). Some of the authors themselves seem to have studied such tests as alternatives to ANOVA, why not including such tests while at the same time advocating for others.**

***See answer to comment 1***

**Comment 10) Type I and II error rates (or probabilities, if theoretical vs. empirical) indeed correspond to alpha and beta. Yet, a type I error is not a probability, although we estimate the probability of a type I error to occur given the model/method and observations. The first paragraph should therefore be rephrased. Although not an excuse, the same error also appear in the thus correctly referenced paper from 2010.**

***We changed accordingly and introduced the following sentences: “A Type I error consists of falsely rejecting the null hypothesis in favour of an alternative hypothesis, and the Type I error rate (α) is the proportion of tests that, when sampling many times from the same population, reject the null hypothesis when there is no true effect in the population. A Type II error consists of incorrectly rejecting the alternative hypothesis in favour of the null hypothesis, and the Type II error rate (β) is the proportion of tests, when sampling many times from the same population, that fail to reject the null hypothesis when there is a true effect. Finally, the statistical power (1-β) is the proportion of tests, when sampling many times from the same population, that correctly reject the null hypothesis when there is a true effect in the population” (lines 173-182).***

**Comment 11) Defining the type II error as "accepting the null hypothesis" means the authors commit to the Neyman & Pearson approach to data testing (as in their 1933 paper), or a more Bayesian mindset (evidence accumulation and for instance likelihood ratios). Yet, experimental psychology usually commits to a mix of Neyman-Pearson and Fisher's approaches, or their mix with NHST (where accepting the null does not exist). A clarification of this choice would be useful for the naive experimental reader (since not coherently adopted throughout the entire paper). The latest review and tutorial by Perezgonzalez (2015) is in my opinion nice to be explicit on these aspects.**

***We totally agree with this comment. There is no question of accepting the null hypothesis, what we meant was failing to reject the null hypothesis. This is now corrected accordingly.***

**Comment 12) A reference to Rasch, Kubinger & Moder (2011) for testing violations of assumptions is provided, but not others from Moder. I was quite surprised, since Moder published several papers on violations in ANOVA, and for instance compared its statistics/performance to Hotellingís T2.**

***Hotelling’s T² apparently very good when n are equal across groups, however one cannot perform it when unbalanced data (Moder, 2010)***

**Comment 13) Testing a parametric assumption using another test statistic is indeed known to be problematic, since we "aim" at accepting the null while measuring divergences or deviations (i.e. from normality or homoscedasticity, or even a t-distribution using a normality test...). While again aiming at accepting the null (thus Neyman-Pearson-ish), the review here does not integrate expectations based on (minimal) effect sizes (on deviations) that may affect the ANOVA test statistic (thus more Fisherian in this sense). This was one of Neyman-Pearson contributions, popularized more than 50 years later by Cohen in psychology. Turning to graphical methods (which is ok and practical in my opinion) simply means dropping the inferencial data-testing approach, and opting for descriptive statistics. A better reviewing of inferencial methods for testing deviations would be useful, or explicitly restricting the content of the paper to focus on statistical software for psychologists (as done in the referenced "Normality Tests for Statistical Analysis: A Guide for Non-Statisticians" for SPSS).**

***In order to focus on our main topic, we removed all considerations on Shapiro-Wilk/KS tests.***

**Comment 14) Again, Delacre et al. (2017) is referenced, but not much earlier papers leading to the same results (including those of Moder, see above). More surprinsingly, the reference itself does not point to any earlier research work on this issue (where there has been decades of both empirical and mathematical work, with the amount of MC simulations simply increasing in recent years).**

***We added a literature review of previous findings: see the section “Consequences of Assumption Violations” (line 170).***

**Comment 15) Also, since the referenced paper (Delacre, 2017) compares t-test and Welch t-test with similar directions, and since the ANOVA is an omnibus test with contrasts coding allowing to project data and test trends/differences, how could we expect not to get the same overall tendancy here?**

***Of course, it is not surprising that the overall tendency is the same. Our main goal was to highlight that our conclusions for the Welch’s t-test could be generalized when there are more than two groups, and that it is even more important to use the W-test instead of the F-test when more than two groups are compared, because the F-test becomes too liberal even when sample sizes are equal across groups (which is not true when one compares only two groups).***

**Comment 16) I can only guess the kurtosis definition given is that of DeCarlo (1997), but non-referenced. Also, see Westfall (2014) for a more statistically correct formulation (the title of the article is self-explaining: "Kurtosis as Peakedness, 1905 ñ 2014. R.I.P.")**

***The given definition was indeed that of DeCarlo (1997). After reading the suggested reference of Westfall (2014), we adapted the definition: “a measure of the tendency to produce extreme values (Westfall, 2014; Wilcox, 2005)” (lines 97-98). However, we still mention that sd and kurtosis are independent (DeCarlo, 1997).***

**Comment 17) RT distributions are right skewed, because the underlying measure is bounded (by 0) and the distribution cannot be normal (as a response distribution to a scale item cannot be normal either). It's thus rather because there is a left-wall and no right-wall, with possibly large RTs (sometimes then filtered out, or the distribution log-transformed to satisfy normality).**

***We modified our explanation taking your example into account: “Third, bounded measure can also explain the non-normal distributions. Such examples can be found in the ﬁeld of neurosciences such as reaction times, that can be very large but never below zero (resulting in right-skewed distributions”* (lines 118-120)**

**Comment 18) I could not make sense of the "One additional source of variability is the presence of unidentified moderators" sentence in the current context. While it is true (although there will always be unidentified moderators, else it would not be a statistical model analysis, but the complex reality), I could not relate it to the between-condition variability studied in this section. I looked into Delacre et al. (2017, p.95-4), which further improved my understanding of the page although being phrased almost identifcally, but which did not contain a reference to Cohen, Cohen, West, & Aiken (2003). The last paragraph on the page was more helpful, but it seems to simply imply randomization was incorrectly performed, or that very important factors to control (or moderators) where left out. If not highly correlated with each others, these factors should lead to the "noise" required to obtain normally distributed residuals, and thus for applying the type of tests considered here, or to slight deviations from homoscedasticity (thus again, the interpretation depends on the minimal effect size of interest for diagnostics). I must have severely misunderstood something here.**

***What we meant is that by definition, a moderator is a variable that will interact with factors, which implies that the effect of the moderator will be different in one condition of the factor than in another condition of the same factor. This could yield heteroscedasticity. We rephrased in order to take it into account: “Since some of these variables can act as moderators, they can generate heteroscedasticity. Indeed, by deﬁnition, a moderator is a variable that will interact with factors, which implies that the eﬀect of the moderator will be diﬀerent in one condition of the factor than in another168 condition of the same factor.” (lines 166-169).***

**Comment 19) There is no step provided for SD-ratio, making it impossible to compute or understand the combinations / simulation plan.**

***We made a more detailed simulation plan. In regard with a literature review, we explained parameters we considered. Considering the SD-ratio, we mentioned “the SD of the population from which was extracted last group was a function of the SD-ratio, with values of 0.5, 1, 2 or 4. The simulations for which the SD-ratio equals 1 are the particular case of homoscedasticity (i.e. equal variances across groups). (lines 289-292).***

**Comment 20) What was in the end the criteria used to determine whether distributions are normal, same for homoscedastic? Hopefully (seeing the "round" numbers), these must be defined a priori (same SD, rnorm or something else) and not based on estimated statistics on simulated samples? From the paper, one cannot know if you implemented all tests/formulae yourselves (to avoid automatic corrections in some R wrappers), or did you use batching based on standard CRAN packages (in which case, which ones)? In this case, and since software evolves, which version was used (for which packages)?**

***As we mention, we manipulated the population parameter values (and not the sample parameter values): “Population parameter values were chosen in order to […] (line 281). The SD of the population from which was extracted last group []… (line 289).***

***We also now mention that we performed Monte Carlo simulations using R (version 3.5.0; lines 273), using the packages and functions that are visible on Github (see the script generate nSim random datasets, performing 3 tests on each dataset, extracting p-values and storing them in a file.R).***

**Comment 21) Since you certainly can rely on statistics (or theoretical derivations from the probability distributions) to estimate deviation from normality of homoscedasticity, why not directly using such measures as continuous predictors? By splitting them into categories, distortions are introduced that affect all statistics (a statistical issue mathematically related to the actual focus of the paper).**

***We do not understand this comment. You seem to belief that we are turning a continuous variable into a categorical variable and thus introducing distortions. Although we agree with this principle, we do not see to what variable you are referring to. If you could clarify we would be happy to bring a more satisfying answer.***

**Comment 22) Why running 1M simulations for each configuration (long and unecessary), while constraining so much the parameter range (and sets of values more generally). For instance, the effect of sample size tends to be exponential, and should be reflected by the chosen values.**

***We could indeed have reduced the number of simulations in order to increase the parameter range. However, or goal was to show tendencies and we believe that our scenarios are various enough in order to achieve this goal.***

**Comment 23) Is normality met (or not) in all conditions, or just in a few (i.e. differences in distributions between-conditions, or between-simulations)? This can be expected from your assumptions/reasons for not considering normality/heteroscedasticity for granted (e.g. reaching a measure bound in one condition more than in another), and yet is not studied here? What "other distributions", generalizations of the normal distribution (skewed normal, t...) or different ones (beta, gamma, exponential, Weilbull...)? is only by reading the code on**[**https://github.com/mdelacre/Welch-ANOVA**](https://github.com/mdelacre/Welch-ANOVA)**that one gets all the information (e.g. on the strategy used to combine distributions and other parameters). Again, justification of how the simulation plan/design was chosen are sparse. Since not relying on natural t-test to ANOVA generalization, which allows to greatly limit the number of combinations required, minimal justification of the choices made would be welcome.**

***The rationale for the design can be found in the supplemental material (p.2 to 6). The choice of tested distributions is now explained in the paper, and visible in all figures. It is now easier to see that all combinations of SD-ratio, n-ratio, SD and sample sizes pairing are tested under all chosen distributions.***

**Comment 24) Although I commit to relying on median/MAD for position/dispersion estimation under weak constraints, non-normality does not seem like the right argument here. Mean/SD are not strongly associated to normal distributions (although we tend to interpret them on normal distributions in psychology for the sake of simplicity). Else, there would be no use to concentration inequalities (such as Chebyshev's inequality).**

***We finally use the mean instead of the median in order to summarize Type I and Type II error rates, for the sake of simplicity. However, the conclusions are not altered by this choice.***

**Comment 25) Figures 1 & 2 are missing from the paper (while tables are present). Since they were the only reason for dichotomizing the continuous statistics on normality/homoscedasticity, this is regretable. I thus hope it simply is a bug on my computer.**

***This has been changed accordingly (we removed the old Figures and introduced new ones).***

**Comment 26) A whole literature is dedicated to the why one should always rely on the maximally powerful test statistic combined with visuals. This is especially true for departure from normality, with entire paper advocating for Shapiro-Wilk. These therefore are quite general and pre-existing recommendations. From a statistical perspective, and with a focus simply on power depending on distribution shape (thus deviations) see Yap & Sim (2011). Comparisons of various types of normality tests.**

***We finally removed the two-step procedure section in order to focus on the robustness (Type I and Type II error rates) of the F-test, F\*-test and W-test.***

**Comment 27) How can balanced design be guaranteed? For many of the same reasons homoscedascity cannot be guaranteed (p.5-7), this is often not the case. Invoked factors in social experiments, attrition (with non missing at random -MAR- data, thus leading to larger deviations from asumptions) are only a few examples.**

***We agree and replace our original sentence by: “Use balanced designs (i.e. same sample size in each group) or sample sizes are similar as possible whenever possible”***

**Comment 28) If balancing was indeed important, why not discussing this factor at all in the introduction. If not an assumption, ok, but then why not also considering distribution changes in shapes across conditions (which is the strongest factor of all, and usually impacts both asumptions which are at the core of the paper).**

***We added clearer simulation plan in the draft, explaining why we also take balancing into account.***

**Comment 29) Liu (2015) is present in the references, but not referenced in text, and yet the title seems quite relevant to the topic (Comparing Welch ANOVA, a Kruskal-Wallis test, and traditional ANOVA in case of heterogeneity of variance), and indeed is when reading through it.**

***We added the reference in the text. In performing a bibliography with tex (using the package papaja in R), we now ensure consistency between reference in the text and bibliography.***  
**Comment 30) Appendix: Given what is missing in the main text relatively to the simulations, the example of application of F / F\* / W could be moved in SOMs. On the contrary, some elements in SOMs are not optional but mandatory to interpret any result.**

***We reworked the supplemental material about results of our simulations. All important results were implemented in the main text, all other comments were removed.***  
  
**Comment 31) SOM: Why adopting K-S test, when concluding S-W should be chosen in the main text?**

***This is not an issue anymore as we removed the section on the two-step procedure.***

**Comment 32: SOM: The text provides some of the lacking information in the paper about the simulations (that should at least clearly point to the adequate SOM sections). Yet, part of this should be included in the main article to make interpretations minimally possible. Also, no discussion or reasons are provided for the different distributions tested.**

***The whole text has been reworked, based on the principle that the main text should be very didactic for social psychologists, and that all technical specificities can be found on github.***

**Comment 32) Github: There is no git description or even readme file to explain how the code is structured. The names would be explicit, this would not be mandatory, but for instance "generate random datasets.R" (on which I focus below) contains code to apply the inferential tests, as well as most of the simulation plan.**

***We tried to improve the git folder, and we also edited a README file.***

**Comment 33) Github: The fact that simulation data are generated and stored one by one by individual calls to the "get\_simu" function (line 163 to 4011) is clearly error prone (and make the simulation non verifiable). If this R file is script generated (I hope so... while I could not find the source), I still don't see the interest but for obfuscation. Although the functions are useful for replication, the calls are not.**

***We removed the calls.***

**Comment 34) Github: The R code structure is also surprising unefficient (for simulations of such scale, or if users would want to extend the simulations, as it was in my case as a reviewer). Individual if/else tests are nested within the "get\_sample" function (where a simple anonymous function or definition of parameters through local environments would have done the job of wrapping the calls to rnorm/rsnorm/...), itself being called within a double-for loop where vectors are incrementally built without pre-allocation... Github: A few function definitions (for the distributions) and a simple expand/apply combination would have done the whole job in a few lines and much less time (even more with packages dedicated to mapping/aggregating on massive amounts of data). This would also allow far greater maintainability, especially extensibility.**

***Thank you for this comment. It was interesting to learn about these packages and we will be very happy to use them in future.***

**Comment 35) Github: These issues are recurrent in all R scripts (with additional dependency to globally defined variables) and I add to look back in git commit history to get a sense of what was used in the paper rather than updated after article submission (although there are deletion/add operations that hide local changes).**

***In order that the reviewers can see how we took their comments into consideration, the draft version that as sent to the IRSP was moved into a Rmarkdown file. Each modification was then tracked on git.***