**07-Apr-2020**

**Dear Dr. Lakens:**

**Thank you for submitting your General Article (AMPPS-19-0062.R1) entitled "Simulation-Based Power-Analysis for Factorial ANOVA Designs" to Advances in Methods and Practices in Psychological Science (AMPPS). The manuscript has now been reviewed, and the reviewer comments appear at the end of this letter. In addition to the original two reviewers, I sent the manuscript to one of our volunteers who agreed to evaluate the usability of tutorials.**

**Both of the reviewers who saw the initial version were pleased with the thoroughness of the revision. I agree that you did an excellent job of addressing the primary concerns they raised. Based on my assessment and these reviews, I am accepting the manuscript pending a minor revision that address the reviewer comments (and my comments below). Most of these comments involve minor clarifications and corrections designed to make it easier for readers to use the R code and Shiny app.**

**There is one more substantive suggestions that I do feel would greatly enhance the usefulness of the code and package. As noted in the previous stage of review, it would be ideal to add an option to calculate the required sample size for a given level of power. The new "usability" reviewer explicitly mentioned that, and it came up in the previous round as well. Many of our readers might want to be able to calculate the sample size needed to achieve a given level of power for an assumed population effect size. That goal might well be the most common one for power analysis. If it is possible to implement that feature without too much trouble, I'd strongly encourage you to do it.**

**I also agree with reviewer 2 that it is important to distinguish population parameters and sample statistics. Doing so could be as simple as adding a "hat" when you're talking about the sample estimate rather than the population parameter (e.g., hat-mu rather than mu).**

**Below, along with the reviewer comments, I've listed a few minor things I noticed when reading the manuscript.**

**When you submit your revision, please include include a letter detailing your point-by-point responses to each reviewer comment and indicating how you changed the manuscript to address them. I will handle the revised manuscript without sending it for further external review. You can submit your revision by logging into https://eur02.safelinks.protection.outlook.com/?url=http%3A%2F%2Fmc.manuscriptcentral.com%2Fampps&amp;data=02%7C01%7C%7C28e01f6d185f4eb58f4008d7db0dda5e%7Ccc7df24760ce4a0f9d75704cf60efc64%7C1%7C0%7C637218724860488886&amp;sdata=%2FANXQrG8fX9I%2B0vjkYsELnQvmUxecp%2BGUI8l9JVJp90%3D&amp;reserved=0 and entering your Author Center, where you will find your manuscript title listed under "Manuscripts with Decisions." Under "Actions," click on "Create a Revision." Your manuscript number has been appended to denote a revision.**

**IMPORTANT: Your original files are available to you when you upload your revised manuscript. Please delete any redundant or outdated files before completing the submission.**

**Once again, thank you for submitting your manuscript to AMPPS and I look forward to receiving your revision. I think this well written tutorial (along with the R code and Shiny app) will provide a useful resource for our readers.**

**Best,**

**Dan**

**--------**

**Daniel J. Simons, Editor**

**Advances in Methods and Practices in Psychological Science (AMPPS) Psychology Department, University of Illinois ampps.editor@gmail.com**

**EDITOR ADDITIONAL COMMENTS**

**• I'd suggest adding a table listing other power analysis packages to the paper as a helpful resource. You could refer to it at the end of paragraph 2.**

**• minor suggestion: name the package SuperPower rather than Superpower to better highlight the "power" part of it.**

**• Typo: "(be entering a “w” or a “b”)."**

**• Typo in shiny screenshot: "specficied"**

**• "which answers the question whether there are any differences between group means" --> technically, it answers the question of whether there are any differences AMONG the group means. "Between" suggests pairwise differences, but the omnibus ANOVA doesn't necessarily suggest any particular pairwise differences. "Among" might lead to less misunderstanding.**

**• It might be good to point out that the power for an N=80 within design is equivalent to an 80\*conditionLevels between design when the correlation is zero (and higher when it is >0. You do the latter, but don't make the point about equivalent power when correlation=0).**

**• typo: "population standard violations varied extremely across conditions" - "standard deviations"**

**Reviewer(s)' Comments to Author:**

**Reviewer: 3**

**Comments to the Author**

**In this manuscript, Lakens & Caldwell introduces an R package and Shiny app that calculates the statistical power for factorial ANOVA as well as t-tests. This is a very useful tutorial as, at least to my knowledge, current alternatives are either unable to handle ANOVA designs with more than 2 factors or are overly convoluted. Even for one-way ANOVAs, many software require additional calculations (such as specific effect sizes) that can vary widely between software, and have confusing defaults (as stated by the authors). Superpower provides an easy and intuitive alternative, especially for ANOVA designs with more than 2 factors. It took me about 2 hours to go through the paper and try out the code.**

**Comments:**

**As the authors pointed out, one of the purposes of carrying out an a-priori power analysis is to ensure that the study is adequately powered, in this case for an ANOVA. I think there should be more explicit mention or even an example on how to obtain the sample size required for a certain power given. The plot\_power function provides one way of doing so, but I feel that explicitly stating that this could be a method for obtaining sample size requirements would be useful especially for readers who are skimming through it. In addition, although the main goal of the paper is for power analysis, I think it would be useful to a wider audience if it could be extended to include sample size calculations.**

**Below are comments more specific to the usability of the tutorial:**

**(a) all necessary materials are available**

**- This might be obvious to most, but I think it would be if the authors explicitly stated the name of the R package (Superpower).**

**- The links to the Shiny apps are not immediately apparent. The authors provide a link to the manual, which contains the links to the apps, but this is not immediately apparent and confused me for a while.**

**(b) any provided code runs successfully**

**- Yes, all provided code ran successfully.**

**(c) step-by-step instructions are clear and unambiguous**

**- For the app, I would have liked more specific instructions for the definition of “Factor & level labels”. I think it is important to mention that the levels of a factor have to be defined before defining the next factor. Perhaps a specific example would be useful. Also, there is a typo (‘young’ instead of ‘yound’)**

**- Perhaps it would be better to have a separate section for the app rather than combining it with the section that first introduces Superpower in R. Or, shift it to the end of the section. It was a little confusing/jumpy when Superpower was introduced in R, then the Shiny app was referenced for a bit, then it jumped back to R. For example, in the text it mentions that “for a visual confirmation of the input a figure is created..” (pg 4, line 4), but this is referring to only the Shiny app although the text preceding and following refers to the R package.**

**The r code can also produce a figure when adding ‘plot = TRUE’, which we now added to the example. Furthermore, we have moved the Shiny app screenshot below, to the location here it is discussed. This should prevent the ‘jumping back and forth’.**

**- Pg 4, line 44: my output does not exactly match the output in the manuscript. Instead of “p\_cheerful\_sad”, I have “p\_condition\_cheerful\_condition\_sad”.**

**True – we shortened these labels in the submission because of layout issues. We can use long labels, and ask the copy editors to make sure it is printed well.**

**- Under the ‘Power for interactions’ section (pg 7), it would be helpful to give an example of how to specify a 3x2 design in R. It is not clear how the means should be specified in a 3x2 design (or for any factorial design with more than 2 factors and 2 levels). I had to refer to the Shiny app to discover that in a 3x2 design, the means are specified in the following manner:**

**a1\_b1, a1\_b2, a2\_b1, a2\_b2, a3\_b1, a3\_b2**

**rather than**

**a1\_b1, a2\_b1, a3\_b1, a1\_b2, a2\_b2, a3\_b2**

**We added a 3x2 design for the new section on the power plot, to add this information on how means are entered to the manuscript.**

**(d) prerequisites specified in the tutorial are accurate and realistic**

**- There was no mention of prerequisites, but I believe that it is obvious to those who use R that the package first needs to be installed and loaded.**

**(e) the tutorial teaches the skill(s) it promises to teach**

**- The tutorial sought to demonstrate how to conduct power analyses, especially for factorial ANOVA designs (for which there are no satisfactory software currently). Two easy-to-use and easy-to-understand solutions were provided in the form on an R package and a Shiny app.**

**That’s good to hear. The preprint has been online for a while, and is downloaded more than 1000 times and is cited 7 times. We are receiving regular feedback from users and will continue to incorporate this in the future.**

**Reviewer: 2**

**Comments to the Author**

**Overall, the authors were responsive to my comments on the previous version of the manuscript. Moreover, as already outlined in my previous review, I think the software described in this paper is useful, mostly because it allows to calculate (or estimate) power for a variety of ANOVA designs as function of population means, standard deviations, and correlations under H1. Thus, except for a few minor issues that are easily corrected (see below), I think the revision is now ready for publication.**

**However, I still doubt that it is a good idea to use the same notation for (population) parameters and (sample) statistics. I am afraid that this will cause confusion and misconceptions in many readers. For example, it might suggest to researchers that it is possible calculate power for an effect size observed in the sample (“observed power”). This of course has nothing to do with power in the statistical sense because power always refers to specific parameters (or population effect sizes) that define H1. However, I leave it to the editor to decide whether different notations for parameters and statistics are necessary.**

**Minor issues:**

**P. 2, left column, 31: delete one “with”**

**P. 2., right column: Figure 1: Cohen’s d is labelled differently in the figure (x-axis) and in the text**

**P. 2., right column, Title of Figure 1: According to the rule specified in Footnote 1, “alternative hypothesis assuming d = 0.5” should read “alternative hypothesis assuming d in the population = 0.5”**

**P. 2., right column, Title of Figure 2: According to the rule specified in Footnote 1, “alternative hypothesis assuming partial eta-squared = 0.0588” should read “alternative hypothesis assuming partial eta-squared in the population = 0.0588”**

**P. 3, left column, 13-15: “The goal of an a-priori power analysis is to determine the sample size to observe a p value smaller than the chosen alpha level WITH A PREDETERMINED PROBABILITY (i.e., THE DESIRED POWER), given an assumption ABOUT the true POPULATION effect size”**

**P. 3, let column, 18: “the sample size” is redundant here, please delete**

**P. 3, left column, 26: According to the rule specified in Footnote 1, “true effect size is” should read “true population effect size is”**

**P. 3, left column, 41: “be” should read “by”**

**P. 5, right column, 39: delete “when”**

**P. 7, right column, Box 2: The notation in Equation (8) differs from the notation in the text.**

**P. 10-11: Many references are incomplete (i.e., lack volume numbers, page numbers, and doi’s)**

**Reviewer: 1**

**Comments to the Author**

**AMPPS-19-0062.R1**

**This is a very good revision that clearly addressed many of my suggestions in the previous review. My focus this round is more on user experience with the Shiny and R package to enhance usability.**

**As a small point of clarification in Box 1, I would suggest referring to the pooled standard deviation rather than the standard deviation as this the most common approach to calculation of Cohen’s d (although Grisson and Kim’s Effect Sizes book does address some limitations of that approach). If you do make explicit reference to the pooled standard deviation, it would add a nice link to ANOVA-type statistics to show that the square root of MSw/in is equivalent to the pooled standard deviation in this case.**

**I ran the included code several times with various permutations. One persistent issue is the speed of processing for the monte carlo power approaches. The paper does note that more simulations take longer but I think that a warning for less advanced users would be useful. I recognized right off the bat that the 100k sims in the first example might take a good bit of time. However, novice users might simply think the code isn’t working. On this note, I tried the example using 100 sims (1.4 seconds), 1000 sims (12.6 seconds), 10,000 sims (2.1 minutes), and 100,000 sims (26.1 minutes). That said, it is likely better to use an example that won’t make people wait for so long to get an outcome. The Shiny app does well to show the progress but if people are using R they may get frustrated.**

**On a related note, some guidance as to how many sims should be used would be helpful. I admittedly, am just guessing here but given norms for bootstrapping (usually around 2,00), I expect that 100k is overkill. In a quick test, I designed around 64 per group. 100k gave power of 80.97, 10k gave power of 80.59, 2k had power of 79.9, whereas exact power came to 80.14. Those are all pretty accurate (if we take exact as the “right” answer here) so I expect that there is not much gained from going bigger.**

**Thanks, this is a very good point. Indeed, the 100k simulations are overkill. We now use 2000 simulations in the example, note this number of simulations should be completed within 30 seconds, and recommend to try out the simulation with up to 1k, and only perform a larger simulation for the final power analysis. We write:**

**“In the code below 1000 simulations are performed, which should take around approximately 15 seconds and yields reasonably accurate results when trying out the power analysis. For most designs increasing the number of simulations to 10000, which would take a few minutes to complete, should give accurate enough results for most practical purposes.”**

**It would be useful to have sample sizes in the output and a multiple reminders about what the sample size refers to. You do note sample size per group but (and I know this from shortcomings of my own package), without explicit reporting in the output regarding what the sample sizes both per groups and overall, users get confused and cite the wrong sample size particularly since the sample size is not reported in the pdf report.**

**It would be useful to be more explicit on the location of the Shiny app up front. The Shiny does work well but there are other versions online and that may lead to some confusion for authors (particularly since this one https://eur02.safelinks.protection.outlook.com/?url=https%3A%2F%2Farcstats.io%2Fshiny%2Fanova-power%2F&amp;data=02%7C01%7C%7C28e01f6d185f4eb58f4008d7db0dda5e%7Ccc7df24760ce4a0f9d75704cf60efc64%7C1%7C0%7C637218724860488886&amp;sdata=KhS8Yq6MZUiqXc4ZhijzLcJMTdYJkxpZquC3cTqyn2o%3D&amp;reserved=0) throws an error.**

**How should we do this? I think we link to the Superpower book, and there on the landing page link to the paper and apps?**

**Chris Aberson**

**Dear Dan,**

**We have worked as much as we could on improvements in our software, and on the revision of the paper. We needed a bit more time, and we apologize for the longer turnaround time. Compared to our earlier submission, our power analysis software has changed in name (it is now called Superpower, compared to ANOVApower), we added the ability to perform power analysis for MANOVA, power analysis for unequal sample sizes, power analysis were the heterogeneity and sphericity assumptions are violated, specify correlations for each pair of variables in repeated measures designs, and the ability to combine variance estimates for individual comparisons (using the emmeans package). We expect the functionality to grow (including planned contrasts, and ANCOVA designs) in the near future, but these developments are not ready for a full release yet. We are quite happy with this progress, and believe this has made the software very competitive in terms of the available options compared to other tools (see below for a comparison).**

**We have also substantially rewritten the manuscript. It more closely follows the tutorial style, with clear examples, demonstration code and screenshots, and less of the mathematical background (which has been moved into 3 boxes, each of which could also be moved to supplementary materials if you desire).**

**Below, we reply to all points raised during the review process. We feel this revision addresses all the issues that were raised. We are very grateful to both reviewers for pointing out two mistakes in the previous submission concerning the underlying mathematical reason for why correlations in within designs increase power, and why F = t2 in 2 group ANOVA designs. Furthermore, we would like to thank the reviewers for the suggested improvements, which we feel have led to a much more educational paper.**

**Below, all non-bold text is from the original reviews in the decision letter. The bold text is our response. Italic bold text contains quotes from the revised article.**

**---**

Both the reviewers and I felt there were a number of things to like about this paper, especially the treatment of power for follow-up analyses (and the actual hypotheses that researchers might want to test). They both raised a number of concerns that need to be addressed in revision. Some of these are fairly major. Consequently, I am inviting you to submit a revised version of the manuscript, but you will need to address all of the concerns they raise (and a couple of them are potential "show stoppers").

Perhaps the biggest concerns are a) the need to justify the use of simulation to calculate power in these cases, and b) the need to demonstrate how this tool goes beyond other power-calculation tools that were not cited.

If simulation isn't necessary for these calculations, that somewhat undermines fit with the special section and the motivation of the core approach used here (a tool providing analytical solutions could still be relevant for AMPPS, though). However, if the provided package doesn't go beyond what other existing ones provide, that would be a more fundamental problem. You need to make a case for the advance you provide over existing packages. Reviewer 2 lists APriot by Lang, 2017. I also was surprised by the lack of mention of PANGEA from Jake Westfall. You need to do a thorough search for other such tools and identify the advantages of yours over theirs. PANGEA, for example, can handle up to 12 factors with any combination of between and within designs and it appears to do so analytically. If your paper doesn't constitute an advance over existing packages, that would limit its appeal for AMPPS.

**Point a): Justify the use of simulation to calculate power**

**The reviews raise the issue if simulation is required in this app, or whether analytic solutions exist for all power calculations our R package allows. First, it is useful to distinguish simulations from analytic solutions. In the case of an analytic solution we can directly calculate the statistical power of a test based on a closed formula. We do not need to generate any simulated data when using analytic solutions.**

**In our manuscript, we write: "*There are two ways to calculate the statistical power of an"ANOVA through simulations. The first is to repeatedly simulate datasets and compute the percentage of statistically significant results. The second is to simulate a dataset that has exactly the desired properties, perform an ANOVA, and use the ANOVA results to compute the statistical power. The first approach is a bit more flexible, but the second approach is much faster. ANOVApower let’s users choose either of these simulation approaches.*"**

**As far as we know, not all computations our app provides have analytical solutions (or if analytic solutions exist, we could not find them). We use two simulation approaches. The most prototypical approach is to perform calculations thousands of times and compute an average. We use this approach in the ANOVA\_power command, which allows users to correct for multiple comparisons. Our app is (again, as far as we know) the only tool that allows a range of approaches to correct for multiple comparisons, and this is not possible analytically. These power analyses are only possible through the slower ANOVA\_power command, which relies on repeatedly simulating data and calculating averages over (preferably) thousands of simulations.**

**It might be useful for other papers in the AMPPS special issue to explain that there is essentially no difference between simulating 100.000 runs, and averaging over them, or creating one perfect simulated dataset that has exactly these averages, and using that dataset to perform calculations. When simulating a dataset that has exactly the desired properties, we can directly compute the power of the test. We expect many papers in your special issue could, in principle, use both techniques – either generate a single perfect dataset, or repeatedly generate data with variation, and average over many trials. We consider both these techniques ‘simulated-based’ power analyses. During the revision process we included additional functionality in the app. Most importantly, we now allow power analyses for unequal variances and violations of sphericity (e.g., Greenhouse Geisser and Huynh-Feldt), and unequal sample sizes per condition. Analytic solutions might exist for these designs, or it might be possible to derive them, but we can quickly perform these power calculations by simulating a single dataset that has exactly the desired properties, performing an ANOVA, and using those results to compute power using the ANOVA\_exact command.**

**It is important to note that the app can also be used for many tests for which analytic solutions exist. This includes ANOVA designs which assume equal variances and do not correct for multiple comparisons (although even there, many software packages, such as G\*power, are limited in the types of designs, e.g., G\*Power does not allow you to calculate power for a 3x3 within subject design).**

**So to conclude, our package relies on simulation in two ways. Whether the second approach (simulating one perfect dataset) makes our paper sufficiently ‘simulation-based’ to be included in the special issue is really based on the scope of the special issue, and thus is up to the editor to decide. We are just as happy to submit our manuscript to a normal issue in AMPPS. The call for simulation papers was a great deadline for us to get this project ready, we believe we have developed a useful tool, and are happy to see it is already used by the community. Where a manuscript introducing this tool is eventually published is less of a concern to us.**

**Point b): Demonstrate how this tool goes beyond other power-calculation tools that were not cited**

**You have asked us to make a case for our package over existing solutions. In this revision we have added an overview of existing solutions in the second paragraph. We also created a table below that gives an overview of a few dimensions power analysis software differs on. In addition to the columns below, we believe having a good package in R is important, as it is becoming the standard software in reproducible workflows. Currently, users have access to pwr, pwr2ppl and Superpower. The pwr package is very limited. pwr2ppl and Superpower currently only differ in coverage. pwr2ppl covers a lot more types of power analyses than ANOVA designs, but Superpower covers a wider range of ANOVA designs than pwr2people, because researchers can specify any design up to three factors of 999 levels. Furthermore, Superpower automatically provides all main effects, interactions, and individual contrasts, as well as adjustments for multiple comparisons. The two packages are very similar, but Superpower focusses more exclusively on ANOVA designs, is also available as an online Shiny app, and provides more functionality.**

**You specifically asked us to compare our package to PANGEA. We have attempted to do this, but as we explored PANGEA, it was not easy to understand what PANGEA actually does. We contacted Jake Westfall, who no longer works in science, and has limited time to help people (which is completely understandable). The code of the app is, as Jake admitted, not annotated at all, making it very difficult for others to build on the code. We asked experts in multi-level modelling for help, and discovered all the experts we knew found it difficult to understand how PANGEA works. The main challenge is the app requires users to specify variance components, but it is not easy to know how to derive these. Since the app is no longer supported (Jake was the only maintainer), and there are no easy to use tutorials that explain how the app should be used, we believe Superpower is a package that will be used more widely than PANGEA for ANOVA designs. PANGEA does multi-level models, which Superpower does not, and we expect researchers will turn to PANGEA (or simulation based solutions) for those designs, but we expect many people will prefer Superpower over PANGEA for ANOVA designs. Given that PANGEA assumes homogeneity, only allows power analysis for equal N in each condition, does not provide individual comparisons by default, does not allow users to plot power across a range of sample sizes, and does not automatically handle multiple comparisons, we believe Superpower is a complementary tool that users will find helpful, even though it does not accommodate multi-level models yet.**

**In the table below, we coded whether power calculations rely on the homogeneity assumption (or whether power can be calculated for unequal standard deviations) whether it allows users to specify individual correlations for within designs (or assumes a single correlation between all pairs), we specify the input that is required, whether the software automatically provides individual comparisons for more complex ANOVA designs, whether the app allows power analysis for ANCOVA designs, allows users to create a plot of the power across sample sizes, whether it provides power analyses for sequential analyses, , whether the app allows you to correct for multiple comparisons, whether the software provides power analyses for linear mixed models, whether the software provides power calculations for unequal sample sizes, and whether the software is available for free. For the two partial entries, pwr2ppl provides functions for a few specific designs, and Russ Lenth's java applet allows users to correct for multiple comparisons in contrasts.**

**Note that the table below is currently not part of our submission - if you feel it would be useful to include it in a supplementary file, we would be happy to do so. We will also include this table in our online manual for Superpower.**

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | HOMOGENEITY ASSUMPTION | SPECIFY CORRELATIONS | INPUT | INDIVIDUAL COMPARISONS | COVARIATES | PLOT ACROSS N | SEQUENTIAL ANALYSIS | ADJUST FOR MULTIPLE COMPARISON | LINEAR MIXED MODEL | UNEQUAL SAMPLE SIZES | FREE |
| APRIOT | NO | YES | RAW DATA | NO | NO | NO | YES | NO | NO | YES | YES |
| G\*Power | YES | NO | EFFECT SIZE | NO | YES | YES | NO | NO | NO | NO | YES |
| MorePower | YES | NO | EFFECT SIZE | NO | NO | NO | NO | NO | NO | NO | YES |
| SAS GLMPOWER | NO | YES | MEANS, SD, R, N | YES | YES | YES | NO | NO | NO | YES | NO |
| pwr | YES | NO | EFFECT SIZE | NO | NO | NO | NO | NO | NO | NO | YES |
| Superpower | NO | YES | MEANS, SD, R, N | YES | NO | YES | NO | YES | NO | YES | YES |
| pwr2ppl | NO | YES | MEANS, SD, R, N | NO | YES | NO | NO | NO | PARTIAL | YES | YES |
| Russ Lenth: Java applet | YES | NO | MEANS, SD, R, N | YES | NO | YES | NO | PARTIAL | NO | NO | YES |
| PANGEA | YES | NO | EFFECT SIZE N, VARIANCE | NO | YES | NO | NO | NO | YES | NO | YES |
| PASS | NO | YES | EFFECT SIZE, OR MEANS, SD, R, N | NO | YES | YES | YES | NO | YES | YES | NO ($395 per year, or $1195) |

The reviewers made a number of other helpful comments, some of them addressing fairly major concerns (and others minor). I won't rehash all of them here, but you should address each of them. Below I highlight a few of them.

• Reviewer 1 suggests walking through a real example in more depth. I agree. For tutorials in AMPPS, the emphasis should be on the how-to aspect - teach users how to use the tool you provide (even with screenshots - Reviewer 2 noted the need for this as well). Tutorials typically are more hands-on and less heavy on the exposition. To make space, I think you can shift some of the background material and "derivation" content to in-detail boxes or the appendix/supplement. You could also put the p-value distribution content into a supplement.

**We removed the paragraphs showing the *p*-value distribution, which was more educational for educational material, than for a tutorial on how to calculate powers. We moved all derivations in dedicated boxes (1 to 3). If you believe the article will be improved if this content is moved to supplements, we can do this in a revision without any problem, but we think this content might clarify power analyses for ANOVA designs for readers who are not too afraid of formula. Most importantly, we added screenshots, and the code as users need to perform the analyses. We also link to the online manual, which is very extensive, and has a lot of detailed examples.**

• I agree with Reviewer 1 that a potentially big contribution of this paper is the discussion of power for more focused hypotheses. People often power for one thing when their hypothesis is about something else. It would be good to expand that discussion (raised in the individual comparisons section). Even if other aspects of your package and App duplicate what's available in other packages, this part might well be a novel contribution

**We have created a dedicated section "Power for individual comparisons" to highlight this useful aspect of our software, and included power estimates for individual comparisons that are low, despite having good power for main effects.**

• Reviewer 2 notes that your package/App don't provide power calculations for multilevel models (hierarchical linear models). I think that's okay for the purposes of this tutorial. Power calculations for such designs is a much bigger challenge, and given the prevalent use of simpler ANOVA designs, I think a tutorial focused on just those aspects is okay for the purposes of a single paper. We likely will have other papers in this special section that deal with power calculation for more complex designs.

**Multi-level power analysis is a challenge. We believe they require a dedicated solution. They require different input, and a dedicated interface. This is so difficult, that even commercial software like SAS has not solved the problem, and PASS has solutions for a few scenarios (and a standalone simulation module). It is well beyond the scope of what we try to accomplish with Superpower. We have seen preprints on this topic, which we think are excellent. If we look at the software that accompanies these preprints, we see no easy way to integrate this in power analysis software for ANOVA designs, further strengthening our belief that these analyses require dedicated software solutions.**

• Reviewer 2 comments on the lack of validation of your power calculations. Do your calculations match those produced by other tools for power calculation? You should provide some evidence that the power calculations you provide are correct. That could be part of a supplement, but you should note it and refer to the documented accuracy of the calculations in the text.

**We had already validated our program against 4 sources upon our first submission (G\*Power, pwr2ppl, code by Brysbaert, and analytic solutions (where these exist)). Throughout the online manual we validate our code against G\*Power and the pwr package in R. In 3 appendices of arcaldwell49.github.io/SuperpowerBook we validate our code against pwr2ppl, MorePower (added during the revision), and code from a recent paper on power analyses by Marc Brysbeart. Finally, for the tests where analytic solutions exist, we have validated our tests against the analytic results.**

**We have included a number of validations as “tests” within the package to ensure that each update passes these validation checks. With the exception of PROC GLMPOWER in SAS, we feel the validation files we have made available are on par or more extensive than other packages in the literature.**

**We have added a sentence in the manuscript that these validation files exist:**

**"*The online manual at http://arcaldwell49.github.io/SuperpowerBook provides detailed examples for power analyses for One-Way ANOVA designs up to three-way interactions for mixed designs, MANOVA analyses, and power analyses for when ANOVA assumptions are violated, as well as examples validating power analyses in Superpower against existing software.*"**

In addition to the reviewer comments, I noted a few minor issues and typos in my reading:

• I think you can cut the two paragraphs at the start of the section on "Calculating power in ANOVA designs." Most of the rest of that section can go into an "in detail" box (it's not absolutely essential for readers to understand the approach you've taken, but it's good to have it in the paper for background).

**We now start with a basic example, where we ask a number of questions that we address later in the example. Then, we briefly repeat the main concepts in power analysis, and show the relation between a two group comparison both as a t-test as an F-test. We believe this is useful to understand some of the points we raise in the basic example.**

**We have moved all formula for effect sizes for ANOVA designs into a dedicated Box (Box 1). This box could be moved to a supplement, if you believe it is too distracting, but we believe it is again useful to more completely understand some of the points we make later in the article**

• "When power is low there is a high probability of concluding there is no effect when the alternative hypothesis is true." --> perhaps be more precise with "alternative hypothesis" here. e.g., even if the true population effect size is...

**Noted and modified to “When power is low there is a high probability of concluding there is no effect when an underlying effect may exist in the population of interest.”**

• When I tested the Shiny APP (Firefox on Mac), it didn't provide a lot of the options discussed in the paper. It showed a t-test option at the top and what looked like a 1-way ANOVA option at the bottom. I'm not sure if it's broken on Firefox or if there are other issues. It would be worth battle-testing it on several platforms.

**We have made a lot of updates to the Shiny apps (see** [**https://arcstats.io/shiny/anova-power/**](https://arcstats.io/shiny/anova-power/)**) and** [**https://arcstats.io/shiny/anova-exact/**](https://arcstats.io/shiny/anova-exact/) **- note that we link to the manual (**[**https://aaroncaldwell.us/SuperpowerBook/**](https://aaroncaldwell.us/SuperpowerBook/)**), and the manual links to the apps, so we can flexibly move around the apps (which require servers, that might move. We also link to an OSF page that is a stable link, and we will update any links to the manual in the future in case this is needed. Although they should have worked before, we have added information and details on how they should be used. We also included a screenshot of the shiny app in the paper (Figure 1).**

• Typos:

**We apologize for the errors, and have changed all within the text.**

- abstract: "researcher can perform" -> "researchers"

- "how changes to the experimental design influences power" --> "influence"

- "each condition enjoy to interact with the voice assistant" --> "enjoy interacting with"

- "and a means of 1" --> "mean"

- "...dark grey areas under the curve marks the..." --> "mark"

- "data is" --> "data are" (throughout)

- Markdown didn't insert f on page 4 of the formatted manuscript and instead showed: "r power\_oneway\_between(design\_result\_2)$Cohen\_f".

Reviewer(s)' Comments to Author:

Reviewer: 1

Comments to the Author

This paper details an R package and Shiny that provide effective and relatively easy-to-implement power analyses for some of the hardest designs to conduct power analyses on. Existing tools like G\*Power can do many of these analyses correctly – however, those tools require a near expert knowledge to ensure correct specification due to confusing naming conventions for different statistics. Additionally, the inclusion of tests for contrasts with adjustments is a nice step forward. That said, I have several suggestions for improving the manuscript.

In reference to equation 8, the paper notes that positive correlations reduce the standard deviations of the difference scores. Practically, this will almost always be true. However, the increase in power comes from the reduction in the standard error of the difference scores (sqrt of se1 squared + se2 squared -2\*r\*se1\*se2. Increases to the effect size caused by the standard deviation of the difference do not occur for all positive correlations – only those that are greater than .50. As an example, given sd=2, a correlation of .50 yields sd diff = 2 whereas a correlation of .1 actually increases the sd diff to 2.8.

**Thanks - indeed, it is not the reduction of the standard deviation of the difference scores that increases power. We appreciate you pointing out this error. We revisited Maxwell & Delaney, 2004, who write: "*The second major potential advantage of the within-subjects design is increased power to detect true treatment effects. Because differences between treatments are obtained by comparing scores within each subject, the influence of the subject main effect has been removed from the error term (see the full model of Equation 22). Thus, systematic individual differences do not contribute to the error term, as they do in between-subjects designs.*" In our first submission, we wrote "A positive correlation reduces the standard deviation of the difference scores and increases the statistical power". This is not correct, but it was correct that any positive correlation increases power. We changed this in:**

**"*A positive correlation reduces the magnitude of the error term by removing systematic individual differences and increases the statistical power.*"**

On the same issue, the paper follows by demonstrating how eta-squared increases (based on correlations) when moving from a between subjects to a within subjects design. However, this change is really more about the metric than a true increase. If we drop down to a two group example with a mean difference of 1 and a standard deviation of 2, we get d = 0.5 for a between groups design. Using the sd diff measures above for a within group design, we’d get d < 0.5 for r<.5, d = 0.50 for r = .50, and d > 0.5 for r > .5. The results are dramatically different for eta-squared (something like .06 vs. .20). Olejnik & Algina (2003) (also Bakeman, 2005) demonstrate that partial eta-squared is not comparable across between and within designs. This lack of comparability makes the choice of eta-squared challenging for the example. The estimates are different because one is an apple and the other an orange. This isn’t to say that it isn’t important to make these points about the correlation and the effect size, only that partial eta-squared is a bad measure for the demonstration as it wouldn’t be the same even when r = 0. Using the Shiny, I simulated effects for 80 total participants with a correlation of .00 and the mean differences and sd noted above. The within subjects partial eta-squared is .11 whereas the between is .06. Perhaps demoing this with a simpler example that uses Cohen’s d would allow you to make the same points but do so using statistical values that provide a clearer comparison. (Note: I certainly recognize the need to address power in terms of partial eta-squared as, despite the value’s limitations, it is what everyone uses).

**Thanks - we agree with the comments by the reviewer that comparing partial eta-squared in a between and within design is difficult. We have therefore removed the direct comparison, and now note:**

**"As explained by Olejnik and Algina (2003) it is difficult to compare η2p across different research designs."**

**This is actually one of the strengths of Superpower, as it requires users to enter the means, sd's, n, and the correlation, which changes much less (or not at all) between designs of different types compared to partial eta squared.**

In discussing G\*Power settings the paper notes a “common mistake” is not clicking “as in SPSS.” I wouldn’t let G\*Power off the hook so easily. Users aren’t making a mistake – G\*Power is by calling something partial eta-squared but using the value in a manner that is completely inconsistent with how everyone uses it.

**We have slightly more strongly noted how the default by Gpower is a peculiar choice, given how all other statistical software calculates partial eta squared in a different manner. We think it is defensible, from a mathematical viewpoint, to have the default G\*Power has, but from a human factors perspective, we think it leads to many errors. We now write:**

**"Box 2 explains how the default calculation of η2p by G\*Power does not depend on the correlation, and therefore differs from how all other statistical software (including SPSS) calculates η2p. This peculiar choice for a default leads to errors for power analyses that include within-subject factors whenever researchers take a η2p reported in the published literature, and enter it in G\*Power as the effect size."**

Throughout the manuscript, I believe that more detailed and richer examples would facilitate more effective application. A real barrier to users effectively carrying out power analyses is not having motivating examples. If the paper could add plausible research situations with a clear discussion of the choice of the patterns of means (e.g., smallest effect of interest) that includes a justification for a certain size and pattern, that would help users get a better sense of how to apply techniques to their own work. There are a lot of papers on power analysis and most either don’t get read or aren’t understood by the majority of readers. Having some examples in the paper that highlight how to conduct an effective power analysis would enhance the value of this contribution and help separate it from the pack.

**We agree about the importance of specific examples. The revision is now less technical, with most of the mathematical information either removed, or moved into boxed. The article starts with the description of some concrete questions a researcher might have. The paper then goes through a set of examples, illustrating (now with code examples) how to perform these analyses in R or the Shiny apps.**

**We have also added the following in the introduction, given that the online manual is incredibly extensive:**

**"*The online manual at http://arcaldwell49.github.io/SuperpowerBook provides detailed examples for power analyses for One-Way ANOVA designs up to three-way interactions for mixed designs, MANOVA analyses, and power analyses for when ANOVA assumptions are violated, as well as examples validating power analyses in Superpower against existing software.*"**

I think the issues in the Power for Individual Comparisons section could be made a bit more forcefully. A huge problem in the power analyses typically included in papers submitted for publication is the reliance on power for omnibus tests without attention to power for the actual hypotheses (e.g., all treatment conditions differ from the control group). An example fleshing that piece out and showing how the omnibus power might be great, but some contrasts could have very limited power would drive that point home.

**We have added a seperate header ("Power for individual comparisons") and now in detail discuss both the benefit of Superpower of providing power for all individual comparisons automatically, and we give an example of low power for individual comparisons, showing that having high power for a main effect is typically not enough. We believe this indeed strengthens the manuscript.**

The p-value distributions section seemed a little thin – this is certainly a useful tool and I’ve seen the code used elsewhere to demonstrate power distributions but, as written, the piece doesn’t seem to link particularly well to the rest of the paper.

**We agree. We have deleted this section from the paper. Storing the plots of p-values is a nice feature, but not important enough to mention it in this tutorial, since the tutorial is focussed on performing a power analysis, and the plot is more useful when educating people about power analysis.**

Finally, it would be helpful to link each demonstration in the paper to specific functions in the package. In several cases (the p value distribution in particular), it took me a bit of exploring to figure out how to use the tools. The readme file on github is very helpful – integrating some of those materials into the paper could help solve this problem. Also, I would have liked to see this piece in the Shiny (as that is where most users will go).

**We have now added code for all the analyses in the tutorial. Again, the online manual also contains a lot of examples, with code, and we hope users will check these out (our clearer link to this online manual in the main text hopefully suffices).**

Chris Aberson

Reviewer: 2

Comments to the Author

Review of AMPPS-19-0062

Content:

This manuscript briefly introduces ANOVApower, an R package to assess the statistical power of various F and t tests for uni- and multifactorial ANOVA designs (both within and between Ss), including follow-up t-tests. Rather than requiring standardized effect sizes as input as most programs do, ANOVApower expects the alternative hypothesis to be defined in terms of means, standard deviations, and correlations between repeated measurements under H1. Statistical power is not calculated analytically but estimated empirically using a Monte Carlo simulation approach.

Evaluation:

This paper is generally well written, clearly structured, and easy to follow. I have not checked the software in detail but it appears to work fine. Moreover, I like the idea to calculate power as a function of the model parameters expected under H1 (specified in terms of populations means, standard deviations, and correlations). This approach is less error-prone than reliance on Cohen’s (1988) effect-size conventions because these conventions have different meanings in different designs. This gives rise to potentially misleading power calculations when users are not fully aware of the definition of effect size measures in different designs and the associated effect size conventions. In sum, this paper might make a nice contribution to the literature.

That said, I also have to point out a number significant weaknesses of both the software and the current manuscript that – in my view – preclude publication of the present manuscript.

Major weaknesses of the program:

1) It is unclear to me why the authors propose a simulation-based approach to estimate power. Use of population means, standard deviations, and correlations as input parameters in no way enforces use of Monte Carlo methods. Of course, power can be calculated analytically because the input parameters map directly onto the noncentrality parameter of the noncentral F distribution. Use of analytical methods would not only be more parsimonious in terms of computation time, it would also be more precise. Why do the authors prefer a computationally very expensive and less precise method over an equivalent method that is computationally relatively inexpensive and more precise?

**First we would like to note that the ANOVA\_exact function is not computationally more expensive than an analytic formula. It takes milliseconds, and is exact (or precise). But the ANOVA\_power function is computationally expensive and less precise. Reasons to use it are to apply sequential adjustments to the alpha level for multiple comparisons (e.g., the holm procedure), or to calculate power when some of the test assumptions are not met (e.g., for within designs).**

**We have added a statement that the ANOVA\_exact approach is generally recommended (unless specific corrections for multiple comparisons are of interest):**

***"The first approach is a bit more flexible (e.g., it allows for sequential corrections for***

***multiple comparisons such as the Holm procedure), but the second approach is much faster (and generally recommended)."***

**We are unaware of any methods that analytically allow the flexibility to calculate power from the means, correlations, and standard deviation from a simple one-way ANOVA to mixed 3-way ANOVA. We believe the current package offers a simple approach to calculating power for a variety of factorial type designs. In addition, as we now note in the manuscript, the simulation based approach allows us to estimate the effect of violating the assumptions of ANOVA on type 1 error rate as well as power - perhaps a somewhat advanced feature that will not be widely used, but it is nice that this is all available in the same software package.**

2) The program is limited to rather simple standard ANOVA models. It cannot handle power calculations for hierarchical linear models and mixed models that have become more important in recent years. Interestingly, the authors cite Brysbaert (2019) as an example for a practical primer of power analyses (p. 1). However, if I didn’t overlook something, their program is not able to assess power for most of the models discussed by Brysbaert (2019).

**We now clearly state in the introduction: "Two limitations of the current version of Superpower is that it does not compute power for ANCOVA or linear mixed model designs."**

**To repeat our response above to the editor:**

**"Multi-level power analysis is a challenge. We believe they require a dedicated solution. They require different input, and a dedicated interface. This is so difficult, that even commercial software like SAS has not solved the problem, and PASS has solutions for a few scenarios (and a standalone simulation module). It is well beyond the scope of what we try to accomplish."**

**We do not see this as a limitation of our software, but as a feature that is well out of the scope of what power analysis software for ANOVA designs tries to accomplish.**

3) ANOVApower offers post-hoc power assessment only (power as a function of model parameters under H1, sample size, and alpha). It does not allow for alternate forms of power analysis that have gained importance in recent years, for example, a priori power analysis (i.e., sample size calculation) and sensitivity analyses (calculation of effect sizes that a given design can detect with sufficient power). These types of power analyses are included in standard power programs like G\*Power.

**It is useful to distinguish 'a-priori' power analysis as 'determining the required sample size based on a desired power, alpha level, and expected effect size' from 'calculating the power, after the effect size, alpha level, and sample size have been specified'. Our app does the first - it allows users to determine the required sample size. The calculation currently requires them to enter, and adjust, a sample size, or they can us the plot-power function and read off the sample size.**

**It would be a rather straightforward extension of the software to allow users to enter the desired power, instead of the sample size, and calculate the required sample size directly. This is a useful suggestion, but in our experience, users understand that they can check the power for different sample size and make an informed decision about the sample size they need.**

**Throughout the manuscript, we state that we are talking about 'a-priori' power analysis. Users can easily vary the means, standard deviation, and correlation to observe the impact of the change of effect sizes on power (e.g., sensitivity analysis). It might be convenient to allow users to enter the desired power and directly return the sample size, and we will consider adding this function in the future.**

Major weaknesses of the paper:

4) The manuscript fails to acknowledge that other software has been published before that uses a simulation-based approach to assess power for various ANOVA designs, including those designs for which ANOVApower was designed. I know of at least one such program (APriot by Lang, 2017). The authors should cite this source (and perhaps others they might identify after a literature search) and discuss the relative advantages and disadvantages of the different programs.

**We have added a discussion in the main text, where we now write:**

***"There is a range of power analysis software available such as G\*Power (Faul, Erdfelder, Lang, & Buchner, 2007), MorePower (Campbell & Thompson, 2012), PANGEA (J. Westfall, 2015a), pwr2ppl (Aberson, 2019), APRIOT (Lang, 2017), PASS, and SAS. These tools differ in their focus (e.g., sequential analyses for APRIOT, linear mixed models for PANGEA), the tests they provide power analyses for (e.g., allowing violations of the homogeneity assumption, unequal sample sizes, and power analysis***

***for ANCOVA designs), and the input they require (e.g., effect sizes, raw data, or means, standard deviations, correlations, and sample sizes). Despite this wide range of software options, in our experience researchers often struggle to perform power analyses for ANOVA designs."***

**and:**

***"Compared to G\*power, the pwr R package, and the pwr2ppl R package, Superpower can compute power for a wider range of designs (e.g., up to 3 factors of 999 level). Compared to PANGEA, G\*power, and MorePower, we believe the required input is somewhat more intuitive, as users enter means, standard deviations, and correlations, instead of effect sizes and variance components."***

**We have included references to other software/programs throughout the paper.**

**In addition, have included extensive documentation of other packages capabilities in comparison to Superpower in a guide book we have written (arcaldwell49.github.io/SuperpowerBook). There are direct comparisons with G\*Power throughout the manuscript. We have in Appendix 1 compared Superpower to pwr2ppl:** [**https://aaroncaldwell.us/SuperpowerBook/appendix-1-direct-comparison-to-pwr2ppl.html**](https://aaroncaldwell.us/SuperpowerBook/appendix-1-direct-comparison-to-pwr2ppl.html)**. pwr2ppl by the reviewer Chris Aberson is one of the few power analysis solutions in R that can do more complex designs.**

5) The manuscript does not distinguish properly between parameters (population level) and statistics (sample level). For example, the authors fail to distinguish between Cohen’s d in the underlying population and Cohen’s d in the sample (e.g. on Page 2, left column). The same applies to f and other parameters/statistics. This is very confusing and potentially misleading (see below). The authors should definitely introduce different symbols for parameters and corresponding statistics and distinguish properly between the population and the sample level throughout their manuscript.

**In past papers we have added a footnote (e.g., Albers & Lakens, 2018: "Note that although these three indices are estimators, we adopt the usual convention to denote them without a ‘hat’"). The difference between population and sample parameters is something statisticians care much more about than users, but where different symbols are considered confusing by our target audience. We agree with the reviewer the distinction should be made explicit, and have done so be adding a footnote that states:**

***"Note that we refer to sample level statistics by default, and explicitly mention whenever we refer to population parameters instead."***

**As a consequence, we have added "population" in 8 places in the manuscript.**

6) P3/4: I missed a systematic description of the program input and output, perhaps illustrated using screenshots. If you do not want to outline this in length in the article, please provide an Appendix or Electronic Supplementary Materials and cite it on Page 3.

**This was also noted by Reviewer 1, and we have added all the code to reproduce the analyses reported in the manuscript. Furthermore, the manuscript is written in RMarkdown, so it is completely reproducible, and all code underlying the reported results can be traced back to the code. Finally, additional examples and screenshots are available in the online manual we now more clearly link to.**

7) P6, right column, 1st par.: You argue against Maxwell et al.’s (2017) recommendation of using t tests in ANOVA designs based on the overall error term (across all groups) because variances might differ between groups. However, if this is actually the case, how can you then use between-groups ANOVA in the first place which is also based on the assumption of homogeneity of variances? Isn’t that inconsistent? I mean, if you trust in the robustness of ANOVA against (small) violations of homoscedasticity, why don’t you trust in the robustness of post-hoc t tests against (small) violations of homoscedasticity?

**This is a fair point, and we leave it up to authors to decide which approach is most appropriate for their design (t-tests vs estimated marginal means). We have added the estimated marginal means power analyses to the software, and to the text, so that those reading the tutorial can compare the results of both approaches. In addition, they could use the ANOVA\_power function to determine the effect of violating the assumption of homogeneity of variance, as we now also note.**

8) Did you perform any checks of ANOVApower against other power analysis tools? If so, please summarize the scope and the outcome of these convergence checks briefly.

**We have included documentation of this within our book arcaldwell49.github.io/SuperpowerBook. As is shown, our package matches the output with G\*power (for the subset of ANOVA designs that can be reproduced in G\*power), Morepower, pwr2ppl, and code by Brysbaert.**

**Further, we have unit tests within the package and so that we ensure future updates to the package do not “break” the package.**

Minor problems:

P2, left column, 2nd par., last line: “there is true effect” –> “there is a true effect”

**Thanks, corrected.**

P2, right column, 2nd par, line 3-5: Although it is true that F = t^2, the reasons given for this equation are misleading and wrong. First, the variance of mean differences is irrelevant for the two-groups t-test. What is relevant is the variance of the means (in addition to the within-groups variance). Second, the variance of the means is not (m1 – m2)^2 but (m1-m)^2, where m denotes the grand mean (provided groups sizes are equal). Again, the paragraph does not properly distinguish between means in the population and means in the sample.

**We are grateful for the reviewer to point out this mistake. In the revision, as we removed mathematical details, the explanation of the relationship that F = t2 has, however, been removed. In Box 1 we still note that F = t2 but we no longer explain it. Nevertheless, we appreciate the reviewer correcting our understanding of the reason for this correspondence.**

P. 3., left column, paragraph below Eq 6: Lambda is the noncentrality parameter of the noncentral F distribution, not a parameter characterizing the distribution of eta-squared under H1 (as illustrated in Figure 2). Please correct accordingly. The easiest way to cope with this problem is probably to replace Figure 1 by the central/noncentral t distribution und Figure 2 by the central/noncentral F distribution.

**In the initial version, we wrote:**

***"Based on λ (which specifies the shape of the expected distribution under the specified alternative hypothesis, e.g., the black curve in Figure 2)*".**

**This language was a bit loose - formally the function for the distribution is df((x\*df2)/(df1-x\*df1), df1, df2, ncp). In our experience teaching statistics it is easier to plot the distribution of Cohen's d and eta-squared, because putting these values on the x-axis is more meaningful for psychologists than the t and F value. To prevent confusion we now write:**

***"Based on λ (which together with the degrees of freedom specifies the shape of the expected distribution under the specified alternative hypothesis, e.g., the black curve in Figure 2)"***

P. 4, left column, last paragraph, 3rd line. Please replace “the statistical power for our design is reduced to 81.14%” by “the statistical power for a 3-groups ANOVA-F test is reduced to 81.14%”. Otherwise it is not fully clear to which test the power value refers.

**We have modified this sentence. It now reads:**

***“the simulations show the statistical power for a 3-groups one-way ANOVA F-test is reduced to 81.14%”***

P. 4, left column, last paragr., line 5: “the mean the mean”: Please delete one “the mean”.

**Apologies for the oversight, the additional mean has been deleted**

P. 4, left column, 8th line from button: Something weird happed here. Please correct the text!

**Again, apologies, we have modified the markdown file so that this error no longer occurs.**

P. 5, 2nd par., line 1: “Revisting” “Revisiting”

**Noted and changed**

Reference:

Lang, A.-G. (2017). Is intermediately inspecting statistical data necessarily a bad research practice? The Quantitative Methods for Psychology, 13(2), 127-140. doi:10.20982/tqmp.13.2.p127