Reply to reviews.

Dear editor, I am very grateful for the three constructive reviews, and your feedback, on my manuscript ‘Sample Size Justification’. I have taken some time to revise the manuscript, both because of other obligations towards collaborators that were more important than a single author paper, as to take some time to see how the preprint was used and cited, and if I needed to make changes based on this. As the preprint is nearing it’s 100th citation, I think the paper is achieving it’s aims rather nicely, but I have made some minor additional clarifications not asked for by reviewers based on how I have seen my paper cited.

Point 1:

Both reviewer 1 and 2, and yourself, note how some worked examples would improve the manuscript. Reviewer 1 notes “One or two examples that demonstrate how to engage with various justifications would go the extra step to provide authors with a template they can use for improving their own articles.” Reviewer 2 says that a chief weakness of the article is “here are a breathtaking number of considerations at play”. The reviewer believes it is “pretty overwhelming to try to perform a power analysis given all these considerations” and predicts that a sample justification will take five days, while using heuristics will take 5 minutes. I can understand this feeling, although the reviewer is exaggerating a little bit here. This paper was written based on 5 years of experience guiding my direct colleagues through sample size justifications, since these became part of our ERB procedure. My colleagues do not need 5 days, but researchers need time to learn how to improve what are typically quite mediocre practices. Raising the bar from the relatively low standards of sample size justifications that are currently the norm in psychology will likely feel a bit overwhelming for many psychologists, and this is exactly the reason that I have kept ‘heuristics’ as a sample size justification. Anyone who can’t, or does not want to, justify their sample sizes can stick to that option, and just honestly admit it.

However, in this revision I have gone the extra mile to make it as easy as possible to provide a solid sample size justification, and I am dedicated to keep facilitating people after this manuscript is published. First, I have created a Shiny app that works as an interactive checklist that guides users through their sample size justification. As the Dutch tax office slogan says ‘We can’t make it any more fun, but we can make it easier’. Second, I have started to collect examples of different approaches to sample size justifications, or aspects of sample size justification (such as justifying an effect size estimate based on an earlier study) and have linked to these from within the Shiny app. I have also added a worked example from my a recent preregistration, where we try to follow best practices as well as possible. Although there are not a lot of best practice examples in the literature, I will monitor articles that cite my paper, and keep collecting examples for users to follow over the next years.

The Shiny app and the worked examples will hopefully make the recommendations feel less overwhelming, and reduce the time it takes to complete a sample size justification.

Point 2: Using heuristics.

You point out that if there is no clear smallest effect size of interest, researchers might use my article to ‘justify’ the use of default levels of Type 1 and 2 error rates (5% and 20%). One of the papers I completed before I returned to this revision deals specifically with 2 practical approaches to set the alpha level for a study to a non-standard level. One approach recommends balancing the costs of Type 1 and Type 2 errors, which is essentially a compromise power analysis based on quantified relative costs of Type 1 and Type 2 error rates. I discuss this approach now. In addition, I directly address the issue of not being able to justify heuristics, and the typical levels, in the section on using heuristics. (I think this is a nice addition – not the sample size uses a heuristic, but the values used to compute a sample size are still often a heuristic).

Minor points.

Reviewer 1 (Chris Abserson) mentions, in addition to a request for a worked example: “For the section on justifying sample size based on resource constraints, I believe the section would be strengthened by including recommendations to provide sensitivity analyses – e.g. “given our sample size, we had power of .80 to detect effects as small as d = .xx, 90% power to detect effects of d = yy, and 95% power to detect d = zz.”

I had written in the section on resource constraints that “From the six ways to evaluate which effect sizes are interesting that will be discussed in the second part of this review, it is useful to consider the smallest effect size that can be statistically significant, the expected width of the confidence interval around the effect size, and to perform sensitivity power analysis.’

However, it is true that a sensitivity analysis is probably the most important tool, so to highlight this more, I have added the following sentence:

**Specific points by Reviewer 2**

Reviewer 2 writes “I would suggest, by comparison, that sample size is less likely to be motivated by concerns regarding error rates than it is a negotiation of sorts between the authors and the journal editor regarding how much sample size is necessary to “earn” a publication”, and adds: “In this light, justifications like “this previous study used this sample size” make sense, as these justifications are not about frequentist error rates, but about a sense of fairness in editors’ expectations regarding publishable sample sizes”. Although I see where the reviewer is coming from, I am not interested in fairness, or expectations about when someone should ‘earn a publication’. As I write in the first sentence of my article the focus in my article is that “Scientists perform empirical studies to collect data that helps to answer a research question.”. Although the questions about whether we should allow sample size justifications based on non-scientific aspects, such as which sample size is ‘enough’ to earn a publication, are very interesting and worthwhile to discuss, they do not fit in the value of information perspective I take in this manuscript. It would be interesting to see a discussion about the factors the reviewer describes in the literature, but I am not the person to write it. Therefore, I have ignored this alternative view on sample size justification in this revision, but I think it would be interesting if someone would write a paper on that topic.

I agree with reviewer 2 that “social psychology has very little power to make urgent or impactful decisions”, and it is difficult for me to imagine a situations where a compromise power analysis would make sense in social psychology (which is the research area I completed my PhD in, so I know it relatively well). Overall, compromise power analyses will be rare, but they will be somewhat more common in applied research projects, for example when people test if it is worthwhile to implement theoretical work in a specific context. I have added that these situations are arguably rare in the section on compromise power analysis.

I have fixed the “imagine if there is be” typo pointed out by Reviewer 2, and added comments to the code snippet for the simulation, as recommended.

**Specific points by Reviewer 3.**

Point 1: Reviewer 3 recommends shortening the article in two places. First, the reviewer recommends to remove the section on sample size justifications in qualitative review. It is an understandable suggestion, but I have decided to keep it in for 2 reasons. First, Reviewer 2 wrote “I think it is neat to see sample size justifications applied to qualitative research.” It is a topic that I feel is important to include (briefly – it is half a page) to show I do not ignore qualitative research, and to highlight the topic also applies to qualitative research (and therefore should not be ignored in those studies). Furthermore, beyond providing some examples of approaches, I believe my criticism “Regrettably, principled approaches to justify sample sizes have not been developed for these alternative philosophies (Marshall, Cardon, Poddar, & Fontenot, 2013).” is also an important point, as there is quite some improvement possible in qualitative research. As the article is ridiculously long anyway, I ultimately feel the last half page hardly matters. Indeed, in the future, I foresee the article will mainly be used as a reference alongside the Shiny App. In other words, I expect users to start in the Shiny app, and jump out to the article to read up on sections that are relevant for their research. In that sense, the length is not very important, as very few people will ever read the paper from start to finish. Those who do will probably read it as part of their education – and in that case, the length is actually quite adequate to accompany a lecture on sample size justification in a research methods course (and somewhat similar to a chapter in a book). Second, the reviewer writes “Another example is p. 10: The author gives several examples for overly simplistic heuristics, where I feel that one example would suffice to make the point.” I agree with this (I got carried away a bit by all the bad heuristics out there, but it is not needed to make the point). I have therefore removed X examples.

Point 2: The reviewer accurately points out I was inconsistent in my discussion about heuristics and ‘valid’ sample size justifications. I want to thank the reviewer for pointing this out. I have addressed the comments raised in the section on heuristics, and in all other sections where heuristics were discussed.

Point 3: The reviewer correctly notes that the idea communicated in Figure 1 is relevant but not communicated clearly. Upon reflection, I agree. I have removed Figure 1, and have rewritten the text to express the core idea in simpler terms.

Point 4: The reviewer believes the use of ‘post-hoc’ power is a confusing term, and suggests using ‘retrospective power analysis’. I like that term, and have added it (and the reference). Searching through my personal Zotero library, I was also surprised to see the term used in over 20 articles I have read and stored – I was not aware it was used so often. However, the term ‘post hoc power” is still used a lot more in my Zotero library (more than 40 articles I have downloaded over the years use it). Therefore, I have decided to follow Lenth (2007), who I believe has written one of the best articles on the topic, who uses ‘post hoc power’ in the title, and in the first sentence immediately adds the second term ‘retrospective power’ to clarify the intended meaning.

Point 5: Thanks for pointing out the inconsistency in the notation of d and delta. Delta is now consistently used to refer to the population effect size, and d to the observed/sample effect size.

Point 6: The author recommends discussing a ‘reverse Bonferroni correction’ and refers to Francis and Thunell (2021). There is a very large literature on this going back decades (e.g., Yekutieli, D., & Benjamini, Y. (1999). Resampling-based false discovery rate controlling multiple test procedures for correlated test statistics. *Journal of Statistical Planning and Inference*, *82*(1), 171–196. <https://doi.org/10.1016/S0378-3758(99)00041-5>). I cannot do this literature justice, but I have formulated the sentence “If researchers are willing to reject a single hypothesis if any of multiple tests yield a significant result, then the alpha level in the a-priori power analysis should be corrected for multiple comparisons” more generally, and now say: “If researchers correct the alpha level when testing multiple hypotheses, the a-priori power analysis should be based on this corrected alpha level.”