Researcher's Intuitions about Power in Psychological Research

Marjan Bakker

Tilburg University

Chris H. J. Hartgerink

Tilburg University

Jelte M. Wicherts

Tilburg University

&

Han L. J. van der Maas

University of Amsterdam

Author Note

Marjan Bakker, Tilburg School of Social and Behavioral Sciences, Tilburg University; Chris H. J. Hartgerink, Tilburg School of Social and Behavioral Sciences, Tilburg University; Jelte M. Wicherts, Tilburg School of Social and Behavioral Sciences, Tilburg University; Han L. J. van der Maas, Department of Psychology, University of Amsterdam.

The preparation of this article was supported by grant numbers 400-08-214 and 452-11-004 from the Netherlands Organization for Scientific Research (NWO).

Correspondence concerning this manuscript should be addressed to Marjan Bakker, Tilburg School of Social and Behavioral Sciences, Department of Methodology and Statistics, PO Box 90153, 5000 LE Tilburg, The Netherlands, telephone: +31 13 466 2964, M.Bakker_1@uvt.nl.

Abstract

Many psychology studies are statistically underpowered (Cohen, 1962, 1994). This may partly be because many researchers rely on intuition, rules of thumb, and prior practice (along with practical considerations) to determine the number of subjects to test. We surveyed 505 published research psychologists regarding statistical power. We found large discrepancies between researchers' reports of their preferred amount of power and the actual power of their studies (calculated based on their typical sample size, effect size, and alpha). Furthermore, 89% of the respondents overestimated the power when asked to estimate the power of specific research designs, and 95% underestimated the sample size to obtain 80% power for studying small effect sizes. Neither experience nor knowledge predicted bias in self-reported power intuitions. Because many respondents based sample sizes on rules of thumb or common practice in the field, we recommend the reporting of formal power analyses.

Keywords: power, survey, methodology, sample size, effect size

Researcher's Intuitions about Power in Psychological Research

Despite the existence of alternative analytical techniques (Rouder, Speckman, Sun, & Morey, 2009; Wagenmakers, Wetzels, Borsboom, & van der Maas, 2011), and notwithstanding criticism (e.g., Nickerson, 2000), null hypothesis significance testing (NHST) remains the main statistical tool in the analysis of psychological research data (Bakker & Wicherts, 2011; Nuijten, Hartgerink, Van Assen, Epskamp, & Wicherts, 2015; Wetzels et al., 2011). Much recent debate on how researchers use NHST in practice concerned the inflation of the number of Type I errors, or rejecting the null hypothesis when it is in fact true (Asendorpf et al., 2013; Bakker, van Dijk, & Wicherts, 2012; Simmons, Nelson, & Simonsohn, 2011; Wagenmakers et al., 2011). To improve the quality of studies and to decrease the possibilities of Type II errors, studies should be well powered (Fiedler, Kutzner, & Krueger, 2012; Simmons et al., 2011).

It has long been argued that researchers should conduct formal power analyses before the start of data collection (Cohen, 1965, 1990), yet many studies in the psychological literature continue to be statistically underpowered (Bakker et al., 2012; Cohen, 1990; Maxwell, 2004). Specifically, in light of the typical effect sizes (ES) and sample sizes seen in the psychological literature, the statistical power of a typical two group between-subjects design is estimated to be less than .50 (Cohen, 1990) or even .35 (Bakker et al., 2012). These low power estimates appear to contradict the finding that over 90% of published studies in the literature have *p*-values below the typical $\alpha = .05$ threshold for significance (Fanelli, 2010; Sterling, Rosenbaum, & Weinkam, 1995). This apparent discrepancy is often attributed to the combination of publication bias (i.e., the non-reporting of non-significant results; Rosenthal, 1979) and the use of Questionable Research Practices (QRPs) in the collection and analysis of data (John, Loewenstein, & Prelec, 2012; Simmons et al., 2011). Despite the centrality of power in NHST (Gigerenzer, 2004), formal power analyses are rarely reported in the literature. Sedlmeier and Gigerenzer (1989) found that none of the 54 articles published in the 1984 volume of the *Journal of Abnormal Psychology* reported the power of their statistical tests. In a more recent and fairly representative sample of 271 psychological papers that involved the use of NHST (Bakker & Wicherts, 2011), only 3% of the authors (explicitly) discussed power as a design consideration of their studies. Thus, it appears that sample size considerations are hardly ever based on formal and explicitly reported (a priori) power considerations.

Here we consider another explanation of the common failure to conduct sufficiently powerful studies, namely intuitions about statistical power. In a classic study, Tversky and Kahneman (1971) showed that even quantitatively oriented psychologists underestimate the randomness in small samples. In addition, Greenwald (1975) asked social psychologists about the acceptable Type II error rate and found an average response of around .27, which means an acceptable power of .73, which again is markedly higher than the overall power computations given by Cohen (1990) and Bakker et al. (2012). These results suggest that researchers may intuitively overestimate the power associated with their own research and that of others (i.e., in their role of reviewers).

Given the centrality of power in the debate regarding reproducibility and replicability in psychology and beyond (e.g., Asendorpf et al., 2013; Button et al., 2013; Gilbert, King, Pettigrew, & Wilson, 2016; Open Science Collaboration, 2015), we conducted two surveys among psychology researchers on their practices, intuitions, and goals in achieving sufficient power. In the first study, respondents assumed roles either as researcher (assessing their own studies), or as reviewer (assessing their peer's studies) and reported how they typically determined their sample size in planning of studies, and what their typical sample size, ES, power, and alpha were. This will inform us about the typical study within the respondent's research line from the two relevant viewpoints as researcher and reviewer, respectively. In the second survey, respondents estimated the actual power and sample size of different (given) research designs.

Study 1

Method

Participants. We collected all email addresses of the corresponding authors of the 1304 articles published in 2012 in *Journal of Consulting and Clinical Psychology, Cognitive Psychology, Developmental Psychology*, *Health Psychology*, *European Journal of Work and Organizational Psychology*, *Cognitive, Affective, & Behavioral Neuroscience*, *Personality and Individual Differences*, *Psychological Methods*, *Journal of Experimental Social Psychology*, or *Psychological Science*. After removing 80 duplicate email addresses and 5 physical addresses, we invited 1219 researchers to participate in our online survey on the Qualtrics website in September 2013. Eighty-four emails bounced, thus we can assume that we were able to contact 1135 researchers from different sub-disciplines in psychology. In our planning of this survey, we expected this sample to be sufficiently large for the (mostly descriptive) analyses that we had planned.

Of all the contacted researchers 499 (44%) started the survey. This might include respondents who started the survey again after we send a reminder. We could not sent a personalized reminder, because we did not want to be able to connect contact information with the given responses. Seven respondents chose not to sign the informed consent and failed to continue with the survey, whereas 291 (26%) respondents finished the survey. Respondents were randomly assigned to complete the reviewer's version or the researcher's version of the questionnaire. One hundred sixty-nine respondents completed the latter version, while 122 respondents completed the former version. For both studies we will use and report the results based on the respondents with complete data. If different results are found while using all available data, we will indicate this below.

Survey. We developed a short survey containing 10 questions (available at https://osf.io/uq97d). The first version contained questions from a researcher's perspective and the second version contained questions from a reviewer's perspective. The last three questions (research field, statistical knowledge, and number of publications) were the same for both versions. We asked the respondents to describe how they generally determined their sample size (for the reviewers: how they assessed the quality of the sample size), and their assessments of typical α, power, ES (in Cohen's *d*), and *N* (cell size) for an independent samples *t*-test. Two additional questions and their results are described in the Supplementary Online Material (SOM). We used all full responses and the design did not involve any additional dependent or independent variables.

Results

Deciding on Sample Size. The first question of the survey asked how researchers generally determine their sample size. A total of 197 respondents answered this open question from a researchers' perspective (this includes answers from respondents who did not finish the survey). Two independent raters scored whether the answers could be assigned to one or more of five different categories. The raters agreed in 93% of the cases (Cohen's kappa = .80). A power analysis was mentioned by 93 (47%) of the respondents (although 20 of them, or 22%, also mentioned practical constraints, like available time and money). Overall, 40 respondents (20%) stated that practical constraints determined their sample size. Furthermore, 45 respondents (23%) mentioned some rule of thumb (e.g., 20 subjects per condition), 41 (21%) based sample sizes on the common practice in their field of research, and 18 respondents (9%) wanted as many participants as possible, to have the highest possible power to detect an effect.

Table 1

Trimmed means of typical (for researchers) and desired (for reviewers) Alpha, Effect Size (ES), *N*, and power given by the respondents, and the power estimates and bias.

	Trimmed mean
α	.05
ES (*d*)	0.39
N (cell size)	34.6
Reported power	0.80
Calculated power (overall)	0.35
Calculated power (based on individual answers)	0.40
Bias	-0.34

Notes: The calculated power (overall) was based on these trimmed means for ES and *N*. The individual power calculations were based on *N*, ES, and α given by individual respondents. The bias was calculated as: (calculated power (individual) – reported power).

The typical study. Respondents in the researcher condition indicated the typical level of α, ES (in Cohen's *d*), *N*, and power in their research, while respondents in the reviewer condition indicated which of these four levels they deemed acceptable as reviewer. Because responses in the researcher condition were very similar to the responses in the reviewer conditions, we present the combined results (see SOM for the separate results by condition). As the distributions were not normal and included outliers (histograms are presented in the SOM), we report the trimmed means (20% trimming) and use robust statistics to increase power and to protect against an unjust estimation of the Type I error rate (Bakker & Wicherts, 2014; Welch, 1938; Wilcox, 2012; Yuen, 1974). Table 1 reports the trimmed means (medians are provided in the SOM). As can be seen, the average typical (or acceptable) ES is *d* = .39, which is somewhat lower than estimates of mean ES based on large-scale meta-analyses in psychology (*d* = .5; Anderson, Lindsay, & Bushman, 1999; Lipsey & Wilson, 1993; Meyer et al., 2001; Richard, Bond, & Stokes-Zoota, 2003; Tett, Meyer, & Roese, 1994). This ES estimate based on meta-analyses is probably an overestimation due to the publication bias often present in meta-analyses (Bakker et al., 2012). Interestingly, the average typical ES found in this study is comparable to the (original) mean ES (*d* = 0.402) of 100 studies in psychology that were recently subjected to replication (Open Science Collaboration, 2015). The average typical cell size of the independent sample *t*-test found in this study equalled 35, which is somewhat higher than the estimates of mean cell sizes found in the literature (20 to 24 participants; Marszalek, Barber, Kohlhart, & Holmes, 2011; Wetzels et al., 2011). Especially for α and power, researchers seem to have a common standard, as 83% of the respondents reported α = .05, and 69% reported power = .80.

In answer to our first question, one respondent indicated that: “I usually aim for 20 - 25 participants per cell of the experimental design, which is typically what it takes to detect a medium effect size with .80 probability”. However, if we calculate power for an independent samples *t*-test with 20 to 25 participants in each condition and *d* = 0.5 (medium ES), the actual power lies between .34 and .41, which is approximately half of the .80 that the respondent mentions that (s)he wants. Considering that 53% of the respondents indicated that they did *not* generally conduct power analyses and 23% of the respondents used some rule of thumb, we wondered whether respondents’ intuitive power analyses were accurate. To investigate whether respondents’ power intuitions were internally consistent, we calculated the power based on the trimmed means reported in Table 1 with the pwr package in R (Champely, 2009). This resulted in a calculated power of .35. Comparably, we calculated the required sample size based on the trimmed means of α, ES, and power as reported in Table 1. This resulted in a required sample size of 105 participants, which is three times as many participants as the trimmed mean of the *N* as indicated by respondents. For each individual respondent, we also calculated the power based on their reported α, ES, and *N*.

Because some of the researchers might have accepted lower power than others, we compared for each respondent the reported power and the calculated power with a robust within subjects t-test. These differed significantly ($t_y(171) = 19.38$, $p < .001$, $\xi = 0.82$, 95% CI = [0.36, 0.44]). We also calculated the bias for each respondent individually by subtracting their reported power from the calculated power based on their *N* and *d*. Eighty percent of the respondents showed a negative bias (recalculation resulted in a lower power that desired), with 33% showing a negative bias larger than 0.5.

Study 2

A majority of the respondents in Study 1 reported a typical power of .80, which is the common standard advised by Cohen (1965) and others. Hence, it might be that our respondents gave the normative answer instead of their typical power even though they might have known that these were not in accordance. The goal of Study 2 was to measure researcher's power intuitions more directly, by presenting examples of research designs with a given α, ES, and *N* to respondents, and asking them to estimate the power of the described research designs. Additionally, we presented examples of research designs with a given α and ES, and asked respondents to estimate the number of participants needed to reach a power of .80.

Method

Participants. We collected all email addresses of the corresponding authors of articles published in 2014 in the same journals as used in Study 1. After removing 1 duplicate email address and 1 email address from a lab member familiar with the hypotheses, we invited 1625 researchers to participate in our online survey on 'statistical intuitions' in February 2015. We did not pursue a formal power analysis because we consider this sample sufficiently large for the purposes of estimation.

Of all the contacted researchers 404 (24.9%) started the survey, with 214 (53.0%) respondents completing the survey. Respondents were randomly assigned to one of the three sample size versions (the small N, medium N, and large N versions). Sixty-seven respondents completed the version that contained a small *N*, 81 respondents completed the medium *N* version, and 66 respondents completed the large *N* version. We will use and report the results based on the respondents with complete data.

Survey. We used a short survey containing 10 questions (available at https://osf.io/ca25h/), in which we asked the respondents to estimate the power of independent sample (2-tailed) *t*-tests in three research situations, where we varied the ES (Cohen's *d* is 0.2, 0.5, and 0.8) and set Alpha on .05 throughout. Respondents were randomly assigned to one of three sample size conditions (total $N = 40$, 80, or 160) to keep the survey short. In three additional questions we asked all respondents to estimate the sample size required for a power of .80 for an independent samples *t*-test and ES of Cohen's $d = 0.2$, 0.5, and 0.8, simultaneously presented with the corresponding correlations (.10, .24, and .37, respectively). Furthermore, we tested respondents' knowledge of power with a single multiple-choice knowledge item (correctly answered by 168, or 78.5%), and asked them to indicate on a 7-point Likert scale how often they conducted a power analysis and to assess their own statistical knowledge on a ten-point scale. Finally, we asked them to indicate their main subfield of psychological research. We used all full responses and the design did not involve any additional dependent or independent variables.

Results

Intuitions about power and sample size. The true power of the research designs presented to the respondents as calculated with the pwr package in R (Champely, 2009) are presented in Table 2 and the lines in Figure 1. Most respondents were not able to estimate these power values well. The 20% trimmed means and confidence intervals of the power estimates given by respondents are given in Table 2 and the dots with CI in Figure 1. Only when d = 0.50 and N = 80, and when d = 0.80 and N = 40, did true power values lie within the 95% CI. The vast majority of respondents (89%) overestimated power for small ES. This is especially worrying given that these are the effect sizes typically found in psychological research (Open Science Collaboration, 2015) and in the estimates given by respondents in Study 1. When ES is large and N > 80, the power of the t-test in this particular design is underestimated.

Table 2

True power and the 20% trimmed means [95% confidence intervals] of the power estimates given by the respondents in the different research situations based on different underlying effect sizes and total sample sizes.

		d = 0.20 (small)	d = 0.50 (medium)	d = 0.80 (large)
N = 40	True power	0.09	0.34	0.69
	Estimated power	0.240 [0.177,0.303]	0.459 [0.414,0.503]	0.660 [0.612,0.709]
N = 80	True power	0.14	0.60	0.94
	Estimated power	0.344 [0.302,0.386]	0.578 [0.534,0.622]	0.768 [0.726,0.811]
N = 160	True power	0.24	0.88	>.99
	Estimated power	0.504 [0.439,0.570]	0.736 [0.690,0.782]	0.876 [0.842,0.909]

A comparable pattern was found when respondents estimated required sample sizes to obtain a power of .80 given an effect with an independent samples t-test. Table 3 gives the true sample size needed in these cases alongside the estimates (20% trimmed means) given by respondents. With a large ES, participants overestimated the required sample size by about 25 participants on average. The mean estimate given by respondents is quite close to the actual value when ES are medium. For small ES, however, the required sample size was underestimated by 95% of the respondents. Whereas respondents estimated 216 participants on average, in actuality 788 participants are needed to obtain sufficient power for such small effects. Given that respondents in the first study indicated that their typical ES to be around *d* = .4 on average, this suggests that researchers typically underestimate required sample sizes needed for studying effects that they deem to be typical. Unexpectedly, we did find a difference in sample size estimates between the three conditions. Respondents in the large sample size condition gave the highest sample size estimates. This might be a carryover effect from the questions about estimating the power in different research situations due to for example anchoring and adjustment (Epley & Gilovich, 2006).

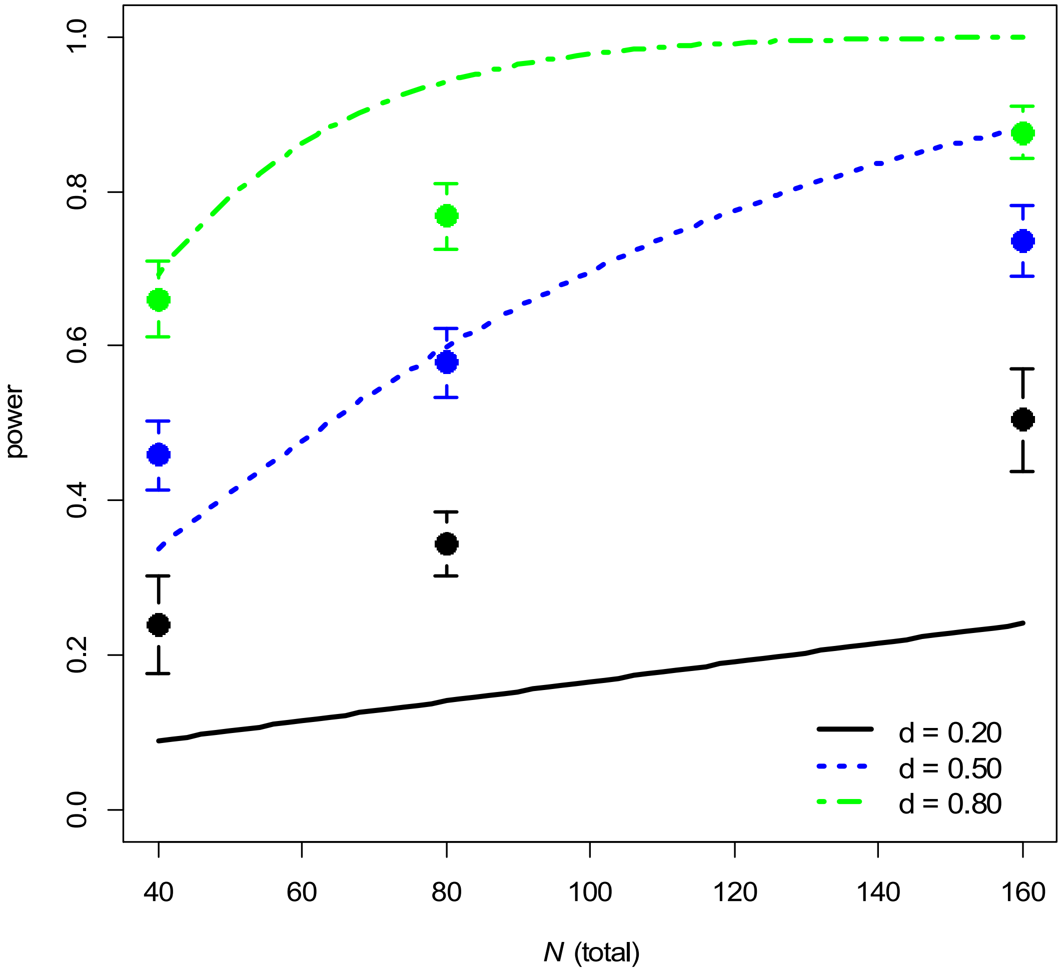

Figure 1: The estimated power (dots) with the 95% confidence intervals. The lines indicate the true power values given the sample size for the three different effect sizes.

Table 3

True sample sizes to reach a power of .8 and the 20% trimmed means [95% confidence intervals] of the sample size estimates given by the respondents for an independent *t*-test and different underlying effect sizes.

	d = 0.20 (small)	*d* = 0.50 (medium)	*d* = 0.80 (large)
True sample size	788	128	52
Estimated sample size (20% trimmed mean)	216 [196,236]	124 [114,134]	77 [72,83]

Other factors (both studies). For both studies, we explored whether some factors might influence the respondents' power intuitions. We focussed especially on the situations with small ES, since these are common in psychology and show the least accurate intuitions. First, those who reported doing a power analysis to determine their sample size did not have better self-reported power estimates. Almost half of the respondents in the researcher's condition in Study 1 indicated that they generally used a power analysis to determine their sample size (although they might not conduct a power analysis for every single study). This group of respondents did not show a significantly higher average calculated power (M_t = .46, 95% CI = [.37, .55]), than the remaining respondents in the researcher's condition who failed to mention the use of power calculations (M_t = .42, 95% CI = [.34, .51]). Furthermore, the amount of bias did not significantly differ between respondents who mentioned typically doing power analyses (M_t = -.31, 95% CI = [-.40, -.22]) and respondents who did not mention typically following that prescription (M_t = -.30, 95% CI = [-.39, -.22]).

Secondly, for Study 2, we also explored the influence of experience and knowledge. Specifically, we summarized the three questions (factual knowledge of power, how often the respondent conducted power analyses, and a self-assessment of statistical knowledge) by means of a PCA. The first component explained 50% of the variance and we used the component scores (CS) to investigate whether these scores predicted estimates of power and sample sizes. Full regression tables and the results of the different questions separately are available in the SOM. We used hierarchical regression analyses. In the first model, we included only CS, in the second model we added condition, and in the third model we added the interaction between CS and condition. The dependent variables were the power and sample size estimates for the research situations with the different ES. For small ES, we did not find any significant effect of CS on power estimates (b = -0.01; t = -0.98, p = .329). Only with medium ES or large ES, did respondents with a high CS have higher (and hence more accurate) power estimates (b = 0.02; t = 2.26, p = .025, and b = 0.04; t = 3.94, p < .001, respectively). Furthermore, when the ES was large, respondents with a high CS had smaller sample size estimates (b = -12.89, t = -2.56, p = .011), again resulting in estimates closer to the true value. We did not find significant effects of CS on sample size estimates when ES was small or medium (b = 16.54; t = 1.33, p = .185, and b = -6.30; t = -1.05, p = .296, respectively). In Study 1 the respondent's self-reported statistical knowledge correlated with neither the calculated power nor the bias, and a robust regression analysis with number of publications yielded no significant prediction.

Lastly, we did not find any significant differences between research fields in estimated power in Study 1 (full results are presented in the SOM). For Study 2, we combined Cognitive Psychology and Neuroscience, and added the two respondents from Forensic Psychology to the other category, because of the low number of respondents in these categories. With a condition (3) by research field (9) two-way ANOVA for trimmed means we investigated the differences between the research fields in Study 2. Only for the situations in which the ES was small, did we find a difference between the research fields in sample size estimates (F_t = 41.43, p = .006). Sample sizes estimates were lowest for respondents from the fields of cognitive psychology and neuroscience and highest for respondents from personality and developmental psychology. However, the highest mean sample size estimate (by the respondents from personality psychology) was 276, which is still far removed from the true required sample size of 788 in this scenario.

Discussion

It has long been noted that the statistical power of studies in the psychological literature is typically too low (Bakker et al., 2012; Cohen, 1990; Maxwell, 2004). The results of the current studies involving over 500 psychology researchers offer insight into why this may be so. Specifically, for most typical ES, respondents overestimated power and consequently underestimated the required sample sizes. When asked about how they normally determined sample sizes in their studies, more than half of our respondents indicated that they did *not* use a power analysis, which may explain why such power analyses are presented in fewer than 3% of psychological articles (Bakker & Wicherts, 2011). Much research in psychology appears to be planned without formal power analysis, and many researchers appear to use rather intuitive approaches in determining their sample sizes.

In our first study we asked researchers and reviewers about their typical study. The actual power estimates based on the reported (typical) sample size and ES were only half of the (typical) power in their research. Over 75% of both researchers and reviewers had a power intuition that resulted in a computed power that was lower than desired. We found similar results from respondents in their role as researcher and in their role as reviewer. In our second study, we asked respondents to estimate the power and sample size given several research designs. We found that 89% of respondents overestimated the power and 95% underestimated sample sizes required for sufficient power when ES are small. The true sample size needed to reach a power of .80 with small ES, is more than three times larger than the mean sample size estimated by the surveyed researchers. This is worrisome, since the results of our first study and replication studies show that ES are often quite small in psychology (Open Science Collaboration, 2015). In combination with publication bias, the (strategic) use of small sample sizes and underpowered studies results in inflated Type I error rates, biased ES estimates, distorted meta-analytical results, and non-replicable findings (Bakker et al., 2012; Open Science Collaboration, 2015).

Even researchers who stated that they typically used formal power analyses showed poor power intuitions. Interestingly, and in line with earlier work showing poor statistical intuitions among mathematical psychologists (Tversky & Kahneman, 1971), self-reported statistical knowledge and experience was not found to be related to better self-reported power intuitions in the most common cases (ES is small). Only when underlying ES is large, did we see some apparent advantage of knowledge and experience. In our second study we found a small difference between research fields in the sample size estimates with smaller ES. However, the true sample size is more than 2.5 times larger than the estimate by respondents from the research field with the highest sample size estimate (personality psychology).

We focused on a between-subjects experimental design, because it is a common and basic design in psychology. Nevertheless, it might be possible that some of our respondents might have been more familiar with other research designs with different associations between sample size and power (e.g., within-subjects design are typically more powerful). However, if experience with research designs had influenced our results, we would expect more differences in power intuitions between sub-fields that involved the use of such different research designs. Future research could focus on power intuitions related to other research designs like within-subjects and correlational designs. We also found some evidence for carryover effects. However, the questions about estimating the power and the sample size showed the same pattern, namely large discrepancies in all conditions when the ES was small. Furthermore, the response rate of both studies was quite low (26% and 13%, respectively), and we expect that researchers who are knowledgeable about power are probably over-represented in this sample due to their interest in the subject. Therefore, we expect that a more balanced sample would lead to an even larger overestimation of power and underestimation of sample sizes in research designs.

Poor intuitions about power may lead to incorrect inferences concerning nonsignificant results. Researchers often conduct multiple small (and therefore likely underpowered) studies of the same underlying phenomenon (Francis, 2014; Hartgerink, Wicherts, & van Assen, 2015). Given the flawed power intuitions, it is quite likely that researchers dismiss these nonsignificant outcomes as due to some methodological flaw (i.e., the notion of a “failed study”) or feel inclined to interpret nonsignificant outcomes as reflecting a true null effect, while in fact these outcomes might be false negatives (Hartgerink et al., 2015; Maxwell, Lau, & Howard, 2015). Therefore, these small (often exploratory rather than confirmatory; Wagenmakers, Wetzels, Borsboom, van der Maas, & Kievit, 2012) studies should be combined within a meta-analysis to estimate a mean effect (and confidence interval) underlying the different studies and to ascertain whether there is heterogeneity in the underlying ES (Bakker et al., 2012).

Our results lead us to the following recommendations when using NHST. First, researchers should always conduct a formal power analysis when planning studies, which is preferably part of IRB approval or preregistration of studies, and they should report this power analysis in their manuscript together with a description of their sample. This will force researchers to explicate their sample size decisions and will likely lead to better-powered studies. Second, considering that often no appropriate ES estimation is available and that our results indicate that intuitions for exponential power functions are often suboptimal and potentially linear, we recommend power analyses to be accompanied by inspecting the implications of offset ES estimates, especially lower estimates. This will help researchers understand the exponential relations involved in statistical power and the considerable impact of seemingly small changes in effect size estimates (see also Perugini, Gallucci, & Costantini, 2014). Third, reviewers should check whether indeed a formal power analysis has been conducted (Asendorpf et al., 2013) and whether it is sound. Fourth, confirmatory studies, or core studies in a research line, should be sufficiently powerful and preregistered (Asendorpf et al., 2013; Wagenmakers et al., 2012). If researchers conduct exploratory studies or analyses, these should be presented as such and possibly combined in a meta-analysis to provide estimates of the mean effect and possible heterogeneity of effects (Bakker et al., 2012).

In the current debate about replicability, reproducibility, and reporting standards, we should keep in mind that researchers and reviewers should collaborate in order to assess the validity of research results (Asendorpf et al., 2013). Both parties may misestimate power of studies, regardless of their self-assessed statistical expertise. There is really only one way: power-up the study.

References

Anderson, C. A., Lindsay, J. J., & Bushman, B. J. (1999). Research in the psychological laboratory. *Current Directions in Psychological Science, 8*, 3-9. doi: 10.1111/1467-8721.00002

Asendorpf, J. B., Conner, M., Fruyt, F. D., Houwer, J. D., Denissen, J. J. A., Fiedler, K., . . . Wicherts, J. M. (2013). Recommendations for increasing replicability in psychology. *European Journal of Personality, 27*, 108-119. doi: 10.1002/per.1919

Bakker, M., van Dijk, A., & Wicherts, J. M. (2012). The rules of the game called psychological science. *Perspectives on Psychological Science, 7*, 543-554. doi: 10.1177/1745691612459060

Bakker, M., & Wicherts, J. M. (2011). The (mis)reporting of statistical results in psychology journals. *Behavior Research Methods, 43*, 666-678. doi: 10.3758/s13428-011-0089-5

Bakker, M., & Wicherts, J. M. (2014). Outlier removal, sum scores, and the inflation of the Type I error rate in independent samples t tests. The power of alternatives and recommendations. *Psychological Methods, 19*, 409-427. doi: 10.1037/met0000014

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munafo, M. R. (2013). Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience, 14*, 1-12. doi: 10.1038/nrn3475

Champely, S. (2009). *pwr: Basic functions for power analysis. R package, version 1.1.1.*

Cohen, J. (1965). Some statistical issues in psychological research. In B. B. Wolmann (Ed.), *Handbook of clinical psychology* (pp. 95-121). New York: McGraw-Hill.

Cohen, J. (1990). Things I have learned (thus far). *American Psychologist, 45*, 1304-1312. doi: 10.1037/0003-066X.45.12.1304

Epley, N., & Gilovich, T. (2006). The anchoring-and-adjustment heuristic: why the adjustments are insufficient. *Psychological Science, 17*, 311-318. doi: 10.1111/j.1467-9280.2006.01704.x

Fanelli, D. (2010). "Positive" results increase down the hierarchy of the sciences. *Plos One, 5*, e10068. doi: 10.1371/journal.pone.0010068

Fiedler, K., Kutzner, F., & Krueger, J. I. (2012). The long way from α-error control to validity proper: Problems with a short-sighted false-positive debate. *Perspectives on Psychological Science, 7*, 661-669. doi: 10.1177/1745691612462587

Francis, G. (2014). The frequency of excess success for articles in Psychological Science. *Psychonomic Bulletin & Review, 21*, 1180-1187. doi: 10.3758/s13423-014-0601-x

Gigerenzer, G. (2004). Mindless statistics. *The Journal of Socio-Economics, 33*, 587-606. doi: 10.1016/j.socec.2004.09.033

Gilbert, D. T., King, G., Pettigrew, S., & Wilson, T. D. (2016). Comment on "Estimating the reproducibility of psychological science". *Science, 351*, 1037-1037. doi: 10.1126/science.aad7243

Greenwald, A. G. (1975). Consequences of prejudice against the null hypothesis. *Psychological Bulletin, 82*, 1-20. doi: 10.1037/h0076157

Hartgerink, C. H. J., Wicherts, J. M., & van Assen, M. A. L. M. (2015). *Too good to be false: Non-significant results revisited.* Retrieved from osf.io/qpfnw

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science, 23*, 524-532. doi: 10.1177/0956797611430953

Lipsey, M. W., & Wilson, D. B. (1993). The efficacy of psychological, educational, and behavioral treatment: Confirmation from meta-analysis. *American Psychologist, 48*, 1181-1209. doi: 10.1037/0003-066X.48.12.1181

Marszalek, J. M., Barber, C., Kohlhart, J., & Holmes, C. B. (2011). Sample size in psychological research over the past 30 years. *Perceptual and Motor Skills, 112*, 331-348. doi: 10.2466/03.11.pms.112.2.331-348

Maxwell, S. E. (2004). The persistence of underpowered studies in psychological research: Causes, consequences, and remedies. *Psychological Methods, 9*, 147-163. doi: 10.1037/1082-989X.9.2.147

Maxwell, S. E., Lau, M. Y., & Howard, G. S. (2015). Is psychology suffering from a replication crisis? What does "failure to replicate" really mean? *American Psychologist, 70*, 487. doi: http://dx.doi.org/10.1037/a0039400

Meyer, G. J., Finn, S. E., Eyde, L. D., Kay, G. G., Moreland, K. L., Dies, R. R., . . . Reed, G. (2001). Psychological testing and psychological assessment: A review of evidence and issues. *American Psychologist, 56*, 128-156. doi: 10.1037/0003-066X.56.2.128

Nickerson, R. S. (2000). Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods, 5*, 241-301. doi: 10.1037//1082-989x.5.2.241

Nuijten, M. B., Hartgerink, C. H. J., Van Assen, M. A. L. M., Epskamp, S., & Wicherts, J. M. (2015). The prevalence of statistical reporting errors in psychology (1985-2013). *Behavior Research Methods*. doi: 10.3758/s13428-015-0664-2

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science, 349*, 943. doi: 10.1126/science.aac4716

Perugini, M., Gallucci, M., & Costantini, G. (2014). Safeguard power as a protection against imprecise power estimates. *Perspectives on Psychological Science, 9*, 319-332. doi: 10.1177/1745691614528519

Richard, F., Bond, C. F., Jr., & Stokes-Zoota, J. J. (2003). One hundred years of social psychology quantitatively described. *Review of General Psychology, 7*, 331-363. doi: 10.1037/1089-2680.7.4.331

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin, 86*, 638-641. doi: 10.1037/0033-2909.86.3.638

Rouder, J. N., Speckman, P. L., Sun, D., & Morey, R. D. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review, 16*, 225-237. doi: 10.3758/PBR.16.2.225

Sedlmeier, P., & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies. *Psychological Bulletin, 105*, 309-316. doi: 10.1037//0033-2909.105.2.309

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science, 22*, 1359 –1366. doi: 10.1177/0956797611417632

Sterling, T. D., Rosenbaum, W. L., & Weinkam, J. J. (1995). Publication decisions revisited: The effect of the outcome of statistical tests on the decision to publish and vice versa. *The American Statistician, 49*, 108-112. doi: 10.1080/00031305.1995.10476125

Tett, R. P., Meyer, J. P., & Roese, N. J. (1994). Applications of meta-analysis: 1987-1992. *International Review of Industrial and Organizational Psychology, 9*, 71-112.

Tversky, A., & Kahneman, D. (1971). Belief in the law of small numbers. *Psychological Bulletin, 76*, 105-110. doi: 10.1037/h0031322

Wagenmakers, E. J., Wetzels, R., Borsboom, D., & van der Maas, H. L. J. (2011). Why psychologists must change the way they analyze their data: The case of psi: Comment on Bem (2011). *Journal of Personality and Social Psychology, 100*, 426-432. doi: 10.1037/a0022790

Wagenmakers, E. J., Wetzels, R., Borsboom, D., van der Maas, H. L. J., & Kievit, R. A. (2012). An agenda for purely confirmatory research. *Perspectives on Psychological Science, 7*, 632-638. doi: 10.1177/1745691612463078

Welch, B. L. (1938). The significance of the difference between two means when the population variances are unequal. *Biometrika, 29*, 350-362.

Wetzels, R., Matzke, D., Lee, M. D., Rouder, J. N., Iverson, G. J., & Wagenmakers, E. J. (2011). Statistical evidence in experimental psychology: An empirical comparison using 855 t tests. *Perspectives on Psychological Science, 6*, 291-298. doi: 10.1177/1745691611406923

Wilcox, R. R. (2012). *Modern statistics for the social and behavioral sciences: A practical introduction.* Boca Raton, FL: CRC Press.

Yuen, K. K. (1974). The two-sampled trimmed t for unequal population variances. . *Biometrika, 61*, 165-170. doi: 10.1093/biomet/61.1.165

Author contribution

M. Bakker developed the study concept. All authors contributed to the study design. Data collection were performed by M. Bakker (Study 1) and C.H.J. Hartgerink (Study 2). M Bakker and J.M. Wicherts performed the data analysis and interpretation. C.H.J. Hartgerink verified all analyses. M. Bakker and J.M. Wicherts drafted the manuscript, and C.H.J. Hartgerink and H.L.J. van der Maas provided critical revisions. All authors approved the final version of the manuscript for submission.

Acknowledgements

We thank Marcel van Assen for his comments on an earlier version of the first questionnaire, and Linda Dominguez Alvarez for rating the first question of the first survey.