# Good science, bad science: Questioning research practices in psychological research

Bakker, M.

Chapter 7

# Flawed intuitions about power in psychological research

Because of relatively small effect and sample sizes, many psychological studies are underpowered. Even if researchers understand the importance of well-powered research designs, their intuitions about power might be incorrect. We surveyed 291 psychological researchers concerning their power intuitions and found large discrepancies between the preferred amount of power and the power calculated based on their typical sample size, effect size, and level of alpha. Furthermore, as reviewers, respondents had even worse power intuitions than as researchers, and preferred multiple small studies over one powerful study. Seniority, sub-field, or self-assessed statistical knowledge were weakly associated with power intuitions. Almost half of the respondents indicated to base sample size decisions on formal power analyses, but the use of power analyses was unrelated to power intuitions. Because many used a rule of thumb or based sample sizes on what is acceptable in the literature, we recommend the reporting of formal power analyses.

This chapter is submitted as: Bakker, M., Wicherts, J. M., van der Maas, H. L. J. Flawed intuitions about power in psychological research.

Despite the existence of alternative analytical techniques (Rouder, Speckman, Sun, & Morey, 2009; Wagenmakers et al., 2011), and notwithstanding the criticism (Cumming, 2013; Nickerson, 2000), null hypothesis significance testing (NHST) remains the main statistical tool in the analysis of psychological research data (Bakker & Wicherts, 2011; 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 et al., 2012; Simmons et al., 2011; Wagenmakers et al., 2011). However, possibilities of Type II errors (failing to reject the null hypothesis when it is false) are equally problematic (Fiedler, Kutzner, & Krueger, 2012; Murayama, Pekrun, & Fiedler, 2013). It has long been argued that researchers should conduct formal power analyses before the start of data collection (Cohen, 1965, 1990, 1994). Yet many studies in the psychological literature continue to be underpowered (Bakker et al., 2012; Cohen, 1990; Maxwell, 2004). Specifically, in light of the typical effect sizes (ES) and sample sizes seen in the literature, typical power 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, 1959; Sterling et al., 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 et al., 2012; Simmons et al., 2011). Despite the centrality of power in the use of NHST, formal power analyses are rarely reported in the literature. Sedlmeier and Gigerenzer (1989) already found that none of the authors of 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 considered (explicitly) the power as a design consideration of the study. Although Bakker and Wicherts found more discussions of power in that sample, these were mostly concerned with explaining non-significant outcomes in the discussion. So it appears that sample size considerations are hardly ever based on formal (and explicitly reported) power considerations.

Here we consider another explanation of the common failure to conduct sufficiently powerful studies, namely flawed intuitions about power (Tversky & Kahneman, 1971). 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 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). This suggests that researchers may intuitively overestimate the power associated with their own research or that of others (i.e., in their role of reviewers). Given the centrality of power in the debate of reproducibility and replicability (Asendorpf et al., 2013; Button et al., 2013), we here survey psychological researchers from different sub-disciplines, either in their role as researcher (assessing their own studies), or in their role as reviewer (assessing their peer's studies) on their practice, intuitions, and goals in achieving sufficient power. Our aim is to answer four questions. First, how accurate are the intuitions of researchers about power? Second, does the role (reviewer vs. researcher) make a difference in assessments of power? Third, are there any differences between sub-disciplines in psychology? Fourth, does (self-assessed) statistical knowledge matter?

The survey can shed light on the accuracy of researchers' intuitions of power in order to explain the discrepancy between the normative notion of power (a formal assessment) and the actual manner in which researchers consider power in their work. In addition, we seek to determine whether these intuitions differ between reviewers and researchers, which is relevant for policy in setting up studies and selecting them for publication.

# 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. Eighty-four emails bounced, thus we can assume that we were able to contact 1135 researchers from different sub-disciplines in psychology.

Our targeted sample size (300) should be able to detect effects of d = 0.32, which lie between small (d = 0.20) and medium (d = 0.50). Of all the contacted researchers 499 (44%) started the survey.[14] Seven chose not to sign the informed consent and failed to continue with

[14] This might contain participants who started the survey again after we sent 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.

the survey, whereas 291 (26%) participants finished the survey.[15] Respondents were randomly assigned to complete the reviewer's version or the researcher's version of the questionnaire. Two hundred sixty-seven respondents started the latter version (169; 63% completed this version), while 225 respondents started the former version (122; 60% completed this version; see also Table 7.1).

Table 7.1
Number of participants per research field and for each condition separately, and the trimmed means of statistical knowledge, recalculated power, and bias per research field.

| | *N* | *N* Researcher | *N* Reviewer | $M_t$ Statistical Knowledge | $M_t$ Power | $M_t$ Bias |
|---|---|---|---|---|---|---|
| Clinical | 43 (15%) | 19 (11%) | 24 (20%) | 6.7 | .47 | -.31 |
| Cognitive | 29 (10%) | 13 (8%) | 16 (13%) | 6.3 | .34 | -.38 |
| Developmental | 42 (14%) | 28 (17%) | 14 (11%) | 7.3 | .40 | -.35 |
| Forensic | 2 (1%) | 1 (1%) | 1 (1%) | 6.5 | .37 | -.03 |
| Health | 12 (4%) | 6 (4%) | 6 (5%) | 7.0 | .47 | -.31 |
| Industrial organizational | 19 (7%) | 11 (7%) | 8 (7%) | 6.8 | .31 | -.45 |
| Neuroscience | 14 (5%) | 10 (6%) | 4 (3%) | 6.6 | .43 | -.31 |
| Personality | 39 (13%) | 22 (13%) | 17 (14%) | 7.3 | .37 | -.32 |
| Quantitative | 10 (3%) | 6 (4%) | 4 (3%) | 8.5 | .44 | -.38 |
| Social | 81 (28%) | 53 (31%) | 28 (23%) | 6.5 | .40 | -.33 |
| Total | 291 | 169 | 122 | | | |

## *Survey*

We developed a short survey containing 10 questions (see Supplementary Materials available at www.bdat.nl). The first version contained questions from a researcher's perspective and the second one 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 participants to describe how they generally determined their sample size (for the reviewers: how they asses the quality of the sample size), whether they would prefer to conduct (or see in manuscript as reviewer) multiple small studies or rather one large study, and their assessments of typical $\alpha$, power, ES (in Cohen's $d$), and $N$ (cell size) for an independent samples $t$ test. Lastly, the survey contained a question about the handling of outliers, which is part of another research project. We used all useful responses and the design did not involve any additional dependent or independent variables.

[15] We will use and report the results based on all the available data. If different results are found for only those respondents who completed the survey, we will indicate this.

# Results

## *Deciding on sample size*

The first question of the survey asked how researchers generally determine their sample size. A total of 197 participants answered this open question from a researchers' perspective. Two independent raters scored whether the answers could be assigned to one or more of five different categories. The raters agreed in 92% of the cases (Cohen's kappa = .80). A power analysis was mentioned by 93 (47%) of the participants (although 20 of them, 22%, also mentioned practical constraints, like available time and money). Overall, 40 participants (20%) stated that practical constraints determined their sample size. Furthermore, 45 participants (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. One participant indicated on the first question 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 an ES of $d = 0.5$ (medium ES), the actual power lies between .34 and .41, which is approximately half of the .80 that the participant mentions that (s)he wants. As more than half of the participants indicated that they did not generally use a power analysis, and 23% of the participants used some rule of thumb, we wondered whether participants were able to perform a good intuitive power analysis.

## *Intuitive power analyses*

Respondents in the researchers 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. As the distributions were not normal and contained outliers, and in line with recommendations by Cumming (2013), we report the trimmed means and use robust statistics to compare the trimmed means (Bakker & Wicherts, in press; Welch, 1938; Wilcox, 2012; Yuen, 1974). In Table 7.2 the trimmed means (20%) are reported for both the respondents who answered these questions from a researcher's perspective (henceforth researchers) and for respondents who answered these questions from a reviewer's perspective (henceforth reviewers). As can be seen, typical (or acceptable) effect sizes center around .4, which is somewhat lower than estimates of mean effect sizes based on large-scale meta-analyses in psychology (Anderson et al., 1999; Lipsey & Wilson, 1993; Meyer et al., 2001; Richard et al., 2003; Tett et al., 1994). The mean desired or typical cell sizes of both

reviewers and researchers in the independent sample *t* test as asked is around 35, which is somewhat higher than the estimates of mean cell sizes found in the literature (Marszalek et al., 2011; Wetzels et al., 2011). Especially for $\alpha$ and power, researchers seem to have a common standard as 83% of the participants reported $\alpha = .05$, and 69% reported power = .80. We did not find significant differences between typical $\alpha$, ES, *N* and reported power between researchers and reviewers.[16]

Table 7.2
Trimmed means for biases in power estimates, and desired (for reviewers) or typical (for researchers) Alpha, Effect Size, *N*, and power given by the respondents

| | Researchers | Reviewers |
|---|---|---|
| Alpha | .05 | .05 |
| Effect size (*d*) | 0.41 | 0.36 |
| *N* (cell size) | 37.3 | 32.2 |
| Reported power | 0.80 | 0.79 |
| Calculated power (overall) | 0.41 | 0.30 |
| Calculated power (based on individual answers) | 0.44 | 0.34 |
| Bias | -0.30 | -0.39 |

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

To investigate whether these power intuitions of researchers were internally consistent, we calculated the power based on these trimmed means with the pwr package in R (Champely, 2009). This resulted in a calculated power of .41 for those who answered questions as researchers and .30 for those who answered as reviewers. We also calculated for individual respondents the power based on their reported $\alpha$, ES, and *N*. The trimmed mean of the calculated power values was significantly higher for participants in the researcher ($M_t = .44$) condition than for participants in the reviewer condition ($M_t = .34$; $t_y(172.7) = 2.49$, $p = .014$, $\xi = 0.23$, 95% CI = [0.02, 0.19]).

Since some of the researchers might have accepted lower power than others, we compared for each participant the reported power and the calculated power. These differed significantly for both the respondents who took a researchers' perspective ($t_y(101) = 11.34$, $p < .001$, $\xi = 0.77$, 95% CI = [0.29, 0.42]) and for respondents who took a reviewers' perspective

[16] The typical $\alpha$ had no variance (after trimming) and therefore the Yuen-test could not be applied; ES: $t_y(159.8) = 1.80$, $p = .074$, $\xi = 0.16$, 95% CI = [-.00, .09]; *N*: $t_y(174.0) = 1.82$, $p = .071$, $\xi = 0.20$, 95% CI = [-.44, 10.63]; power: $t_y(73) = 1.46$, $p = .15$, $\xi = 0.10$, 95% CI = [-.00, .02].

($t_y(72) = 16.40, p < .001, \xi = 0.88$, 95% CI = [0.40, 0.51]). We also calculated for each researcher individually the bias, or the calculated power minus the reported power. The trimmed mean of the bias was smaller for the participants in the researcher condition ($M_t$ = -.30) than for the participants in the reviewer condition ($M_t$ = -.39), although this difference was not significant ($t_y(168.6) = 1.97, p = .051, \xi = 0.17$, 95% CI = [-.0.00, 0.17]). Of those who responded from a researcher's perspective, 76% showed a negative bias (recalculation resulted in a lower power that desired) and 31% a bias of more than -0.5. In the reviewer condition, 86% of the respondents displayed a negative bias, while 34% of respondents showed a bias in excess of -0.5.

## *Doing power analyses and intuitions about power*

Conducting regular power analyses may improve intuitions about power. Almost half of the participants in the researcher's condition indicated that they generally used a power analysis to determine their sample size, though they might not conduct a power analysis for every single study. However, this group of respondents did not show a significantly higher average calculated power ($Mt$ = .46), than the remaining respondents in the researcher's condition who failed to mention the use of power calculations ($Mt$ = .43; $ty(100.0) = 0.51, p = .611, \xi = 0.07$, 95% CI = [-.09, .15]). Interestingly, the amount of bias did not differ between participants who mentioned typically doing power analyses ($Mt$ = -.31) and participants who did not mention typically following that prescription ($Mt$ = -.30; $ty(99.3) = .15, p = .885, \xi = 0.04$, 95% CI = [-.13, .11]).

## *Differences between reviewers and researchers*

Although we did not find significant differences between respondents who answered the questions from a researcher's perspective and respondents answering questions from a reviewer's perspective in $\alpha$, ES, $N$, and power, we did see a difference in the calculated power. This is possible because the participants in the reviewer condition reported a lower (albeit not significantly different) ES and $N$ than those in the researcher condition.

We also saw a difference in answers on how reviewers and researchers would allocate resources in setting up studies. Based on arguments described by Bakker et al. (2012), we know that the use of multiple small samples represents the most optimal choice if one's goal is to find at least one significant effect, but that the use of one large sample represents the optimal choice if the goal is to gather accurate estimates of an effect (if indeed there is a single effect). In that sense, one would expect that if presented with the decision of how to allocate resources for 100 participants, researchers will be tempted to run 5 small studies ($N$ = 20 each) and perhaps not report all of them. On the other hand, in this scenario, reviewers would choose a single large

study ($N$ = 100) because it yields the more accurate estimate of the effect. Interestingly, we found that differences between the conditions in whether respondents would prefer 5 studies ($N$ = 20), 4 studies ($N$ = 25), 2 studies ($N$ = 50) or 1 study ($N$ =100; see Table 7.3) in precisely the opposite direction. A 2 (researcher v. researcher) by 4 (number of studies) $\chi^2$ test was significant ($\chi^2(3) = 25.6$, $p < .001$, $\varphi = .26$). A majority of the participants who answered the question from a researcher's perspective preferred one large study, whereas most participants who answered the question from a reviewer's perspective preferred two smaller studies. So in their role as reviewers, the respondents appeared to prefer two studies with half of the number of participants and thus a lower power above one large study with more power. So in our survey, reviewers opted for less powerful studies than did researchers, despite the fact that smaller studies lead to more inflated type I error rates and effects sizes due to publication bias and/or QRPs (Bakker et al., 2012).

Table 7.3
Number of researchers (%) that preferred 5 studies ($N$ = 20), 4 studies ($N$ = 25), 2 studies ($N$ = 50) or 1 study ($N$ = 100) per condition.

| | Researchers | Reviewers |
|---|---|---|
| 5 studies ($N$ = 20) | 10 (4%) | 14 (9%) |
| 4 studies ($N$ = 25) | 18 (8%) | 13 (8%) |
| 2 studies ($N$ = 50) | 63 (28%) | 80 (49%) |
| 1 study ($N$ = 100) | 134 (60%) | 57 (35%) |
| Total | 225 | 164 |

## *Statistical knowledge*

To see whether respondent's self-assessed statistical knowledge was related to better power intuitions, we correlated[17] the calculated power and bias with respondent's self-reported statistical knowledge. In both conditions, we failed to find significant correlations (Power, Researcher condition: $r_s$ = -0.01, $p$ = .865; Power, Reviewer condition: $r_s$ = .03, $p$ = .763; Bias, Researcher condition: $r_s$ = -0.11, $p$ = .144; Bias, Reviewer condition: $r_s$ = -0.10, $p$ = .265).

## *Differences between research fields and number of publications*

We investigated possible differences in power intuitions between the different research fields. In Table 7.1 the trimmed mean of the calculated power and bias are presented for each research field separately. Because only two participants indicated Forensic psychology as their

[17] We used Spearmans Rank Order correlations, because the data are not normally distributed.

main field of research, we could not include them in a two way ANOVA of trimmed means.[18] We did not find a main effect of research field ($F_t$ = 7.32, $p$ = .622) or an interaction between field and condition in estimated power ($F_t$ = 15.26, $p$ = .155). Similarly, bias showed neither a main effect for research field ($F_t$ = 3.01, $p$ = .951), nor an interaction between field and condition ($F_t$ = 7.69, $p$ = .580).[19]

Table 7.4
Number of participants (%) in each number of publication category and the trimmed mean of the calculated power and bias per category.

| Number of publications | *N* | $M_t$ Power | $M_t$ Bias |
|---|---|---|---|
| < 5 | 41 (14%) | 0.35 | -0.38 |
| 5-15 | 69 (24%) | 0.36 | -0.33 |
| 16-30 | 83 (29%) | 0.40 | -0.35 |
| 31-50 | 39 (13%) | 0.42 | -0.36 |
| 51-100 | 31 (11%) | 0.51 | -0.25 |
| > 100 | 28 (10%) | 0.44 | -0.35 |

We also investigated whether the number of publications of respondents was related to their power intuitions. In Table 7.4 the trimmed means of the calculated power and bias are presented for the different publication categories. A robust regression, by using the rlm() function of the MASS package in R, with condition and number of publication as predictors failed to show a significant main effect of number of publications or an interaction between the research output and condition. [20]

[18] We used the function `t2way()` of the WRS package, which does not give the *df* or ES. Furthermore, we find slightly different results for the main effects of condition compared with the results of the robust *t* test that we used before, because the means are trimmed in every cell (9*2) and for the *t* test in only two cells.

[19] We found no differences between sub-fields or interactions with condition for the reported $\alpha$ and reported power (both no variance after trimming). For ES we did not find a main effect of sub-field ($F_t$ = 16.95, $p$ = .113), but did find a significant interaction effect with condition ($F_t$ = 23.18, $p$ = .032). The estimated ES differed between the conditions for participants whose main field of research was Health Psychology, Personality Psychology, and Social Psychology with an estimated ES for participants in the researcher condition of $M_t$ = 0.29, $M_t$ = 0.40, and $M_t$ = 0.41, respectively, and for the participants in the reviewer condition of $M_t$ = 0.46, $M_t$ = 0.27, and $M_t$ = 0.30, respectively. We also found a main effect of sub-field on $N$ ($F_t$ = 21.44, $p$ = .032), but no interaction effect with condition ($F_t$ = 17.31, $p$ = .081). Especially, the trimmed mean of participants from Clinical Psychology ($M_t$ = 44.6) or Personality Psychology ($M_t$ = 47.1) showed higher values of the reported $N$ that participants from Cognitive Psychology ($M_t$ = 27.1) or Neuroscience ($M_t$ = 26.3).

[20] The number of publications or the interaction with condition did not predict significantly α, ES, and power. However, number of publication positively predicted $N$ ($p$ = .002).

# Discussion

It has long been noted that the power of studies in the psychological literature is typically too low (Bakker et al., 2012; Cohen, 1990; Maxwell, 2004). The results of the current study involving over 250 psychological researchers offers insight on why this may be so. 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, and 23% of the participants used some rule of thumb. This is in line with the finding that such power analyses are presented in fewer than 3% of psychological articles. Much research in psychology appears to be planned without formal power analysis and so many researchers appear to use rather intuitive approaches in determining their sample sizes, and these intuitions are quite far off the mark. Even researchers who stated that they typically use a formal power analysis did not have better power intuitions than those who did not. Intuitions by both reviewers and researchers about (typical) power are effectively twice as large as the actual power estimates based on the reported (typical) sample size and ES. More than 75% of both researchers and reviewers had a power intuition that resulted in a computed power that was lower than desired, and among more than 30% of our respondents this bias exceeded -.50. Interestingly, this tendency to overestimate power was larger for those who assessed results in their role of reviewers, which is somewhat counterintuitive because combined with publication bias and QRPs low power results in severe inflations of the Type I error rate and effect sizes (Bakker et al., 2012). These are problems one would expect reviewers to be more concerned about than researchers who present a significant outcome on the basis of an underpowered study.

Also relevant for the role of reviewers in assessing power of studies was our finding that reviewers prefer the work under review to involve two smaller studies (with lower power) rather than one larger (more powerful) study. So our results may help explain why underpowered studies continue to be used so widely in the psychological literature; researchers and reviewers may strongly overestimate the power and this overestimation of power may be even larger when reviewing other researchers' work. Despite their important role in heightening replicability and reproducibility of results (Asendorpf et al., 2013), reviewers may not have the best power intuitions needed for their task. In fact, many high impact journals nowadays require multiple replications, but such papers often show incredible results due to an excess of significant findings in light of the low power of the reported studies (Francis, 2012b; Schimmack, 2012).

The tendency to overestimate power was not appreciably different across different subfields in psychology. Interestingly, and in line with earlier work showing poor statistical

intuitions among mathematical psychologists (Tversky & Kahneman, 1971), self-assessed statistical knowledge was not found to be related to the tendencies to overestimate power of typical studies. Moreover, a larger number of publications, as a measure of seniority and research experience, did not appear to render researchers immune to flawed intuitions about power of the typical studies in their field.

A majority of the participants reported a typical power of 0.8, which is the common standard advised already by Cohen (1965). It might therefore be that our respondents gave the normative answer, instead of their typical power, even though they might have known that these are not the same. Nevertheless, in our results, the mean difference between the general norm and the power based on the typical experimental design was rather large. To measure researcher's power intuitions more directly, we might in future work present examples of research designs with a given $\alpha$, ES, and $N$ to researchers, and ask them to estimate the power of the described research designs.

We focused on a between-subjects experimental design, because it is a common and basic research design in psychology. Nevertheless, it might be possible that some of our respondents had no or limited experience with this design so that answering our questions might have been out of scope for these participants (e.g., some participants indicated that they worked primarily with preexisting datasets and did not have to make sample size decisions). Other participants might have been more familiar with other research designs with different associations between sample size and power (e.g., a within-subjects design with the same number of participants will result in more power). However, if experience with research designs had influenced our results, we would expect more differences in power intuitions between sub-fields that involve the use of such different research designs. In future research, we could investigate the power intuitions of researchers about other research designs like within-subjects and correlational designs. Another problem might be that some participants were not familiar with interpreting or reporting effect sizes in Cohen's $d$. However, almost all respondents gave realistic answers. In future research we might provide researchers the possibility to answer this question with their preferred effect size measure and subsequently convert them to one common effect size measure. Our study might also suffer from selection bias, since 26% of the invited researchers finished the survey. However, we expect that this bias would lead to an overrepresentation of interested researchers who are well aware of power issues. Such selection bias is not expected to lead to an exaggeration of the poor power intuitions we documented.

Poor intuitions about power may lead to incorrect inferences concerning non-significant results. Researchers often conduct multiple small (and therefore likely underpowered) studies of the same underlying phenomenon. Given the flawed power intuitions, it is quite likely

that researchers feel inclined to interpret non-significant outcomes as reflecting a true null effect (i.e., the notion of a "failed study") while in fact these outcomes are simply due to chance and not statistically distinguishable from results from the other studies that did show significant outcomes. Some non-significant findings are even to be expected under these circumstances (Francis, 2012b; Schimmack, 2012). Therefore, these small (often exploratory rather than confirmatory; Wagenmakers et al., 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 effect sizes (Bakker et al., 2012).

Our results lead us to the following recommendations for the use of NHST. First, researchers should always conduct a formal power analysis when planning studies, which is preferably part of IRB approval or preregistration of studies. This will hopefully lead to better-powered studies. Second, researchers should always report this power analysis in their manuscript together with a description of their sample. Third, reviewers should check whether indeed a formal power analysis was conducted (Asendorpf et al., 2013) and whether it is sound. Particularly manuscripts that present solely significant results on the basis of relatively small samples should be scrutinized for the possibility of an excess of significant outcomes (Francis, 2012b). 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 main study of a research line and be more open concerning exploratory studies.