InvestigationOfSampleSizePlanningInPsycSci

Felix Singleton Thorn

30 September 2018

In order to get an initial estimate of the research planning practices common in psychology I assessed the 121 empirical research articles published in the November 2017 to August 2018 issues of Psychological Science. The sample size was determined in order to constrain multinomial confidence interval width to a maximum of 20%.

Of the 121 empirical research articles published during this period 51 articles reported a power analysis, 42% of sampled articles, Wilson score interval [34%, 51%]. Of the reported power analyses, the most common approach was to effect size selection was to use a Single previous study as the effect size, with 12 articles ( 10 % of articles) reporting having done so.

Despite the fact that pilot studies are almost by definition too small to reliably estimate the true population parameter value of interest, 3 studies reported having estimated the effect size with this value 2% of articles.

Almost as many used benchmarks from Cohen (1988; n = 9, 7 % of articles). Other articles either reported a sensitivity analysis (showing the effect size that the sample size gave them 80% power to detect, n = 6 (5% of articles) in order to justify the obtained sample size. 7, 6% of articles, did not provide any justification for the effect size they reported having used in power analysis, and 4 articles (3% of all articles surveyed) did not state the effect size that they used in a reported power analysis.

Just 3, articles, 2% of the examined articles, reported that they adjusted their estimates for publication bias, and all of these articles used ad-hoc methods such as doubling the sample size that resulted from a power analysis or using the lowest reported effect for an intervention as opposed to the more sophisticated methods that have been proposed.

Table [1] The number and percentage of papers reporting each type of justification for the effect sizes reported in their power analysis.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Effect size selection method | n | Percent | 95\_CI\_LB | 95\_CI\_UB |
| 8 | No power analysis reported | 70 | 58 | 50 | 67 |
| 12 | Single previous study | 12 | 10 | 2 | 19 |
| 4 | Informal assertion of effect size | 7 | 6 | 0 | 15 |
| 6 | Medium effect benchmark from Cohen | 6 | 5 | 0 | 14 |
| 11 | Sensitivity analysis | 6 | 5 | 0 | 14 |
| 7 | No effect size stated | 4 | 3 | 0 | 12 |
| 1 | Average effect size in a set of studies (not a formal meta-analysis) | 3 | 2 | 0 | 11 |
| 9 | Pilot study | 3 | 2 | 0 | 11 |
| 13 | Small effect benchmark from Cohen | 3 | 2 | 0 | 11 |
| 2 | Average effect size in a set of studies (not a formal meta-analysis), reduced for publication bias | 2 | 2 | 0 | 11 |
| 3 | Effect size from meta-analysis | 2 | 2 | 0 | 11 |
| 5 | Lowest effect size reported in a previous paper on this topic | 1 | 1 | 0 | 10 |
| 10 | Rule of thumb supported by power analysis | 1 | 1 | 0 | 10 |
| 14 | Smallest effect size from set of pilots | 1 | 1 | 0 | 10 |

The sample is not representative of psychology articles in general, but does give an….

|  |  |
| --- | --- |
| Category | Example |
| Average effect size in a set of studies (not a formal meta-analysis) | Given the effect sizes of prior work on theories of intelligence (average d = 0.65; Dyczewski & Markman, 2012; Miele & Molden, 2010; Rattan, Good, & Dweck, 2012) and help seeking (average d = 0.60; Bohns & Flynn, 2010; Flynn & Lake, 2008; Nadler & Chernyak-Hai, 2014; Nadler, Harpaz-Gorodeisky, & Ben-David, 2009), an a priori power analysis indicated that about 40 participants per condition would provide sufficient power (greater than 80%) to detect group differences tested at a false positive rate of 5%. |
| Average effect size in a set of studies (not a formal meta-analysis), reduced for publication bias | To determine sufficiently powered sample sizes in the following studies, we reviewed effect sizes of several previous tests of experimentally manipulated pathogen cues on responses to appearance-relevant stimuli (Ackerman et al., 2009; Miller & Maner, 2012; White et al., 2013). We observed an average d of 0.65. To compensate for potential effect-size inflation, we estimated a d of 0.45, which is closer to effect sizes obtained by field-wide meta-analyses (Richard, Bond, & Stokes-Zoota, 2003). This suggested that a sample size of 158 would be needed to detect an effect with 80% power. |
| Effect size from meta-analysis | We chose a task typical of those used to measure frequency discrimination in studies of dyslexic readers (23 studies consisting of 554 control and 582 readingdisabled participants; mean effect size: Cohen’s d = 0.7; Hämäläinen et al., 2012). The samples of children with dyslexia, language impairment, and typically developing controls were large enough (at least 64) to detect an effect of this size with a power of 0.8. |
| Informal assertion of effect size | Using G\*Power (Version 3.1.9.2; Faul, Erdfelder, Lang, & Buchner, 2007), we estimated that with a total sample of 41 participants, we would have 80% power to detect a medium effect, dz = 0.45, given a two-tailed test and an α level of .05. Therefore, we targeted a sample size of approximately 40 participants, with final numbers reflecting the number of students who sought research-participation credit during the semester. |
| Lowest effect size reported in a previous paper on this topic | Given that Bastian et al. (2013) showed effect sizes (fs) from 0.40 to 0.73, we calculated our sample size on the basis of an estimated effect size of 0.40 (a large effect size). This required a sample size of approximately 150 participants for the study to achieve 99% power. We thus aimed for approximately 50 participants per condition (also consistent with the recommendations of Simmons, Nelson, & Simonsohn, 2013). Given the exclusion criteria developed prior to conducting the study, 21 participants who failed to follow instructions to write an essay or failed the attention check were excluded from analyses. The final sample thus consisted of 156 participants (68.6% male; age: M = 30.1 years, SD = 10.1; 49 in the unethical condition, 54 in the ethical condition, and 53 in the neutral condition). |
| Medium effect benchmark from Cohen | To test this, we recruited participants via Amazon Mechanical Turk to play a single round of a one-shot, anonymous HDG. Past research has shown that the scale of competition has a medium to very large effect size (Barclay & Stoller, 2014; Barker & Barclay, 2016; West et al., 0 10 20 30 40 50 60 70 80 90 100 0 20 40 60 80 100 Expected Losses per Pair Inequality (v) –100 –80 –60 –40 –20 0 20 40 60 80 100 0 20 40 60 80 100 Expected Wealth per Pair Inequality (v) a b Fig. 3. Theoretical effects of inequality (v) and the scale of competition (a) on expected wealth and losses. We assume that q = q\* and k = 100. Curves represent a = 0 (dotted curves), a = 1/3 (dashed curves), and a = 2/3 (solid curves). The left graph displays the expected wealth per pair in the model (v – kq2 ). It shows that pairs earn less as competition becomes increasingly local, because local competition increases the frequency of Hawks and therefore the chance of a costly fight. The right graph displays the expected losses per pair in the model (kq2 ). It shows that more wealth is lost as inequality grows and that this effect is more severe as competition becomes increasingly local. 828 Krupp, Cook 2006); using a Cohen’s d set conservatively to 0.5 and statistical power set to .95, analysis suggested a minimum of 88 participants per condition under a nullhypothesis-significance-testing approach (which we did not use). |
| No effect size stated | We recruited 210 participants via Amazon Mechanical Turk in order to obtain sufficient statistical power (> .80) to detect small-to-medium-sized effects in a between-participants design with two conditions. |
| No power analysis reported | No power analysis reported |
| Pilot study | In a pilot study (see the Supplemental Material), this correlation (r) had a value of −.36. Assuming this effect size, we needed a sample size of 76 to achieve 90% power; we ran additional participants to maintain sufficient power after necessary participant exclusions. Informed consent was obtained from all participants. All procedures were approved by the Stanford University Research Compliance Office. |
| Rule of thumb supported by power analysis | Sample sizes were predefined as 45 subjects who did not meet any of the above exclusion criteria (following Simonsohn, 2015), which was 2.5 times the original sample size.1 Power analysis using the G\*Power package (Faul, Erdfelder, Buchner, & Lang, 2009) and the original effect size of Mudrik and Koch (2013) showed that this sample size would give over 99% power to detect an effect. |
| Sensitivity analysis | Assuming a medium effect size (R2 ) of .13, setting power at .80, a sample size of 89 with six predictors would be sufficient to obtain a significant effect at α = .05. |
| Single previous study | The sample size was determined by (a) referring to previous research to obtain a medium effect size (f 2 = ~0.2 in Experiment 1 of the study of Shen et al., 2016) and (b) a power analysis, in which, by setting the alpha level at .01, the suggested sample size was approximately 20 individuals to reach a medium effect size (f 2 = 0.15, according to Cohen, 1988) given our experimental design. |
| Small effect benchmark from Cohen | Given that our primary focus was on the behavior in the real-life dilemma, we wanted to ensure that we had sufficient power to detect small effects on the rate of consequentialist versus deontological judgment in this group, and we tuned our sample size accordingly. We calculated that a sample of 200 participants would have enough power to detect a small effect, OR = 1.68 (equivalent to a Cohen’s d of 0.20; Chen, Cohen, & Chen, 2010), assuming that the distribution of the consequentialist versus deontological decisions in our real-life dilemma would not be extremely unbalanced. In particular, a sample of 200 would have 75% to 95% power to detect small effects at incidence ratios from 50:50 (equal distribution of the alternatives) to 85:15 (strongly unequal distribution). |
| Smallest effect size from set of pilots | In our own 3 replications of the inaction-effect using scales rather than dichotomous options (reported in a different submission) we obtained the following effect sizes: Cohen d = .51, .81, and .38. We based our sample on the most conservative effect of .38, although we expectedthat in this scenario the effect is likely to be much stronger. Using G\*Power alpha = .05, one-tail (direction of hypothesis known), d = 0.38 and power .95 we require a sample of 151 in each condition, 302 overall.The final sample was 316 |