A Powerful Nudge? Presenting Calculable Consequences of Underpowered Research Shifts Incentives Towards Adequately Powered Designs

Supporting Information

Table of Contents

1000 Words

Calculations

This section outlines how we used power information to calculate expected productivity, false positive rates, and replication rates. We also will post R code for performing these calculations on WG's website upon publication.

Productivity

Given the following parameters:

k	Number of studies
p	Probability null is false
$(1-\beta)$	Power
α	Type I error rate

The number of statistically significant effects generated breaks down as follows.

$$Significant\ Findings = [\{(p)\ (1-\beta)\}\ +\ \{(1-p)\ (\alpha)\}]\ k$$

When the null hypothesis is false (on the left, at probability p), the probability if a significant result is the study's power. When the null hypothesis is true (on the right), the probability of a significant result is α. These probabilities are then multiplied by the number of studies run.

Holding p to 50%, and α to .05, we can calculate expected number of significant results for Researcher A and Researcher B. Given the same number of total participants for each, if B runs k studies, A can run $4k$ studies. We can plug in power estimates using R's pwr package.

A's productivity

$$A\ productivity = [\{(.5)\ (.28)\}\ +\ \{(.5)\ (.05)\}]\ 4k$$

B's productivity

$$B\ productivity = [\{(.5)\ (.80)\}\ +\ \{(.5)\ (.05)\}]\ k$$

On any given study, A has a ~16.5% chance of obtaining a significant result. B has a ~42.5% chance. But A runs four times as many studies. Thus, A is around 56% more productive (depending on rounding error throughout).

% Type I Errors Among Significant Results

Using the same basic setup, we can figure out the expected % of each researcher's significant results from Type I errors. Call this the false positive rate among significant results, $\not f$. It is equivalent to calculations for positive predicted value (see Button, et al., 2013), only calculates the proportion of false positives, rather than the proportion of true positives.

If the likelihood of getting a significant result on any study boils down to:

$$\{(p)\,(1-\beta)\} + \{(1-p)\,(\alpha)\}$$

then the false positive rate is simply the ratio of significant results from Type I errors over the total number of significant results.

$$\not f = \frac{\{(1-p)\,(\alpha)\}}{\{(p)\,(1-\beta)\} + \{(1-p)\,(\alpha)\}}$$

Subbing in values gives us:

$$\not f_A = \frac{\{(.5)\,(.05)\}}{\{(.5)\,(.28)\} + \{(.5)\,(.05)\}} = .15$$

$$\not f_B = \frac{\{(.5)\,(.05)\}}{\{(.5)\,(.8)\} + \{(.5)\,(.05)\}} = .058$$

Replication Rates

There are two ways to replicate a study:

1) The initial null hypothesis was true, the initial result was a Type I error, and the replication attempt was also a Type I error.
2) The initial null hypothesis was false, the initial result was not a Type II error, and the replication attempt was not a Type II error.

To simplify this, if the initial significant result was a false positive, replications will be successful at α. If the initial significant result was not a Type I error, replication success will depend on power.

$$Prob\ Replication = \not f\alpha + (1-\not f)\,(1-\beta)$$

Subbing in values, we have:

$$Prob\ Replication_A = (.15).05 + (.85)\,(.28) = .245$$

$$Prob\ Replication_B = (.06).05 + (.94)\,(.80) = .755$$

Describing Self-Reported Research Practices

Participants provided self-report estimates of their 1) typical effect sizes, 2) typical per-condition sample sizes, and 3) their percentage of correct hypotheses. These three values enable calculations of expected power, Type I errors, and replication. These values are summarized below. Although 178 participants completed our focal measure and most demographics, more than 40 participants declined to provide any information about their own practices. Many participants provided ranges for their estimates (e.g., d = .3-.5). We coded both the lowest and highest portion of the range as separate variables.

Table S1. Self-reported research practices

	N	*M*	*SD*	*Min*	*Med*	*Max*
Effect Size Low	135	.39	.13	.1	.4	.8
Effect Size High	135	.4	.16	.2	.4	1
N/Condition Low	134	49.14	40.76	15	40	400
N/Condition High	134	57.43	58.22	20	50	500
Correct Low	131	.58	.17	.2	.6	1
Correct High	131	.59	.16	.2	.6	1

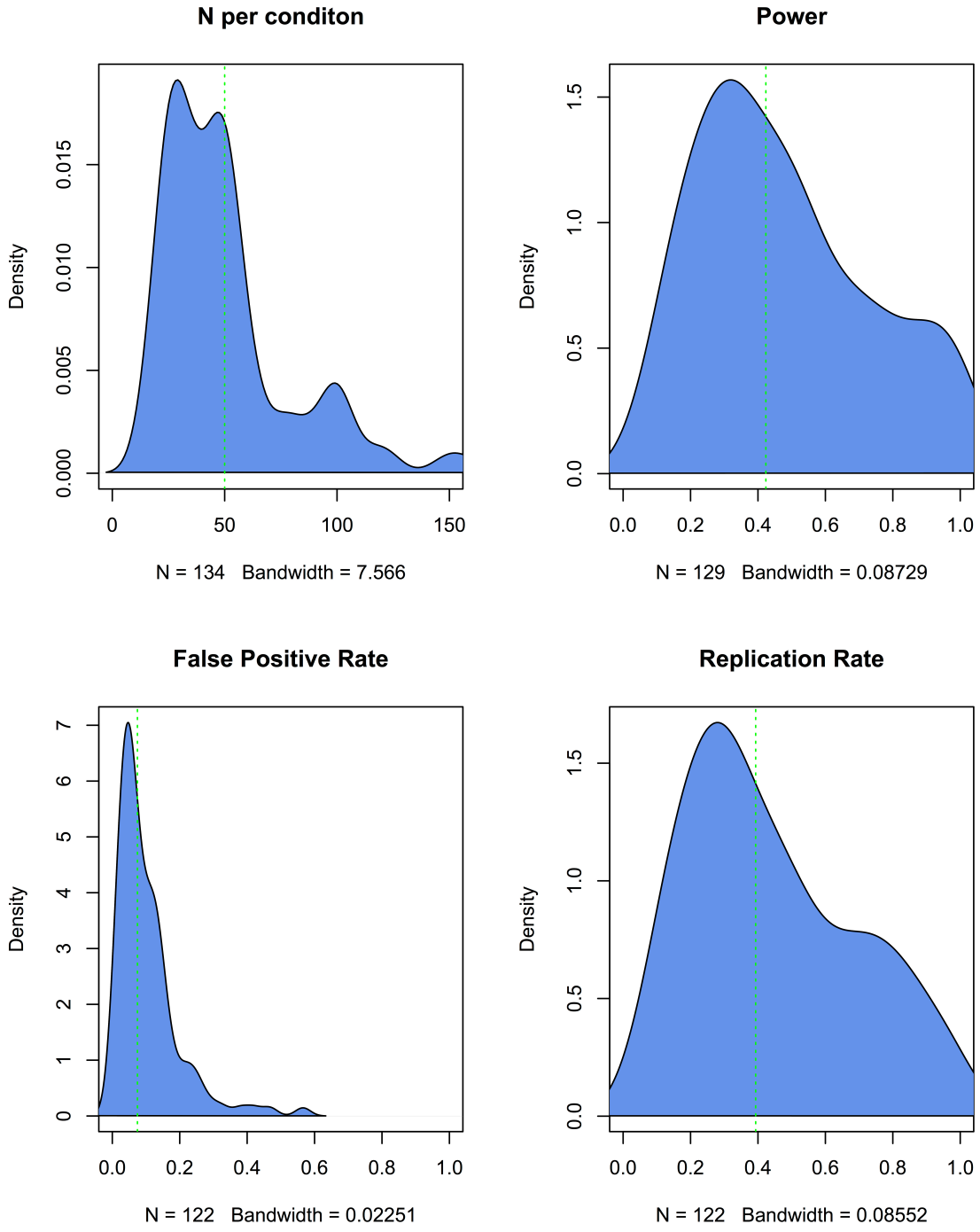
Based on these reported practices, we calculated each participant's power, Type I error rates, and expected replication rates (1N and 2.5N). To obtain conservative estimates, we used only the highest estimate provided by each individual participant.

Table S2. Calculated consequences of self-reported research practices

	N	*M*	*SD*	*Min*	*Med*	*Max*
Power	129	.48	.26	.09	.42	1
Type I Error Rate	122	.10	.09	.00	.07	.57
1N Replication Rate	122	.44	.25	.07	.39	.99
2.5N Replication Rate	122	.69	.24	.10	.73	.99

Figure S1 summarizes the full distributions of four of these outcomes.

Figure S1. Kernel density plots of per-condition sample sizes, power, false positive rates, and 1N replication rates.



Potential Moderators

In addition to confirmatory tests of our focal preregistered hypotheses, we also performed several exploratory tests to explore potential moderating variables. This included tests of whether preferences differed across academic demographics (rank, institution type). In addition, we tested whether reported practices had any predictive power for preferences across conditions.

Did preferences differ across academic demographics?

We performed chi-square tests to explore whether preferences in any condition differed across academic demographics.

Table S3. Summary of academic demographic moderation tests across conditions.

	df	χ^2	*p*
Rank			
Findings	4	3.42	.49
Sample Size	4	6.19	.19
Consequences	3	.82	.84
Institution			
Findings	4	3.88	.42
Sample Size	3	3.27	.35
Consequences	4	1.04	.90

Did preferences differ across reported research practices?

Next, we conducted logistic regression analyses to see whether 1) self-reported per-condition sample size, and 2) calculated power exert any predictive effects on preferences across conditions. Because only two conditions actually included sample size information, we focused only on the Sample Size and Findings conditions. For the logistic regressions, we scaled sample size in units of 10 and power in units of .1. Odds ratios should be interpreted accordingly. Higher values correspond to a greater likelihood of selecting the large sample size candidate.

Table S4. Logistic regression summaries of research practices predicting preferences

	Odds Ratio	**Low CI (2.5%)**	**High CI (97.5%)**	***p***
N per condition				
Sample Size	1.14	.94	1.48	.26
Consequences	1.65	.93	7.87	.35
Power				
Sample Size	1.71	.81	4.05	.18
Consequences	1.21	.78	1.94	.39