Statistical Power: Power's Role in the Replication Crisis, and Justifying Your Sample Size

Roger Strong
Harvard University

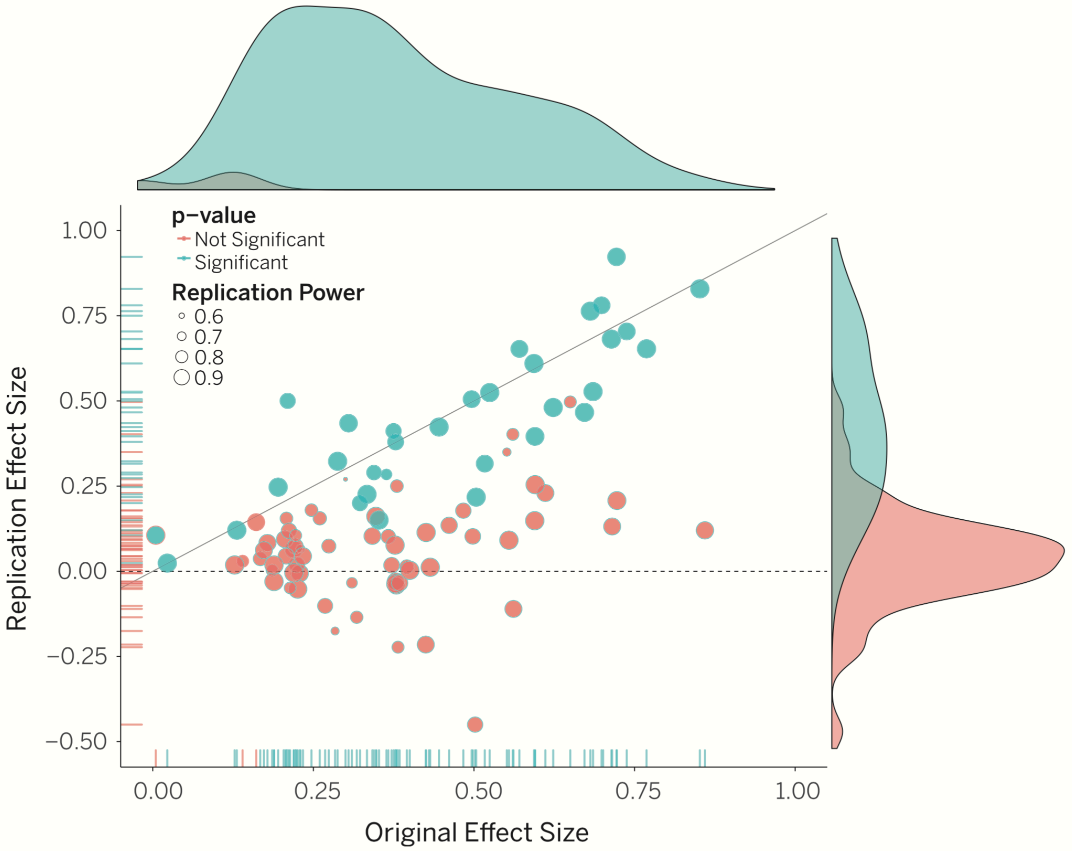
HARVARD
UNIVERSITY

Vision Sciences
Laboratory

Replication Crisis: many studies don’t replicate (red dots)

- one reason: underpowered studies

- Figure from: http://dx.doi.org/10.1126/science.aac4716



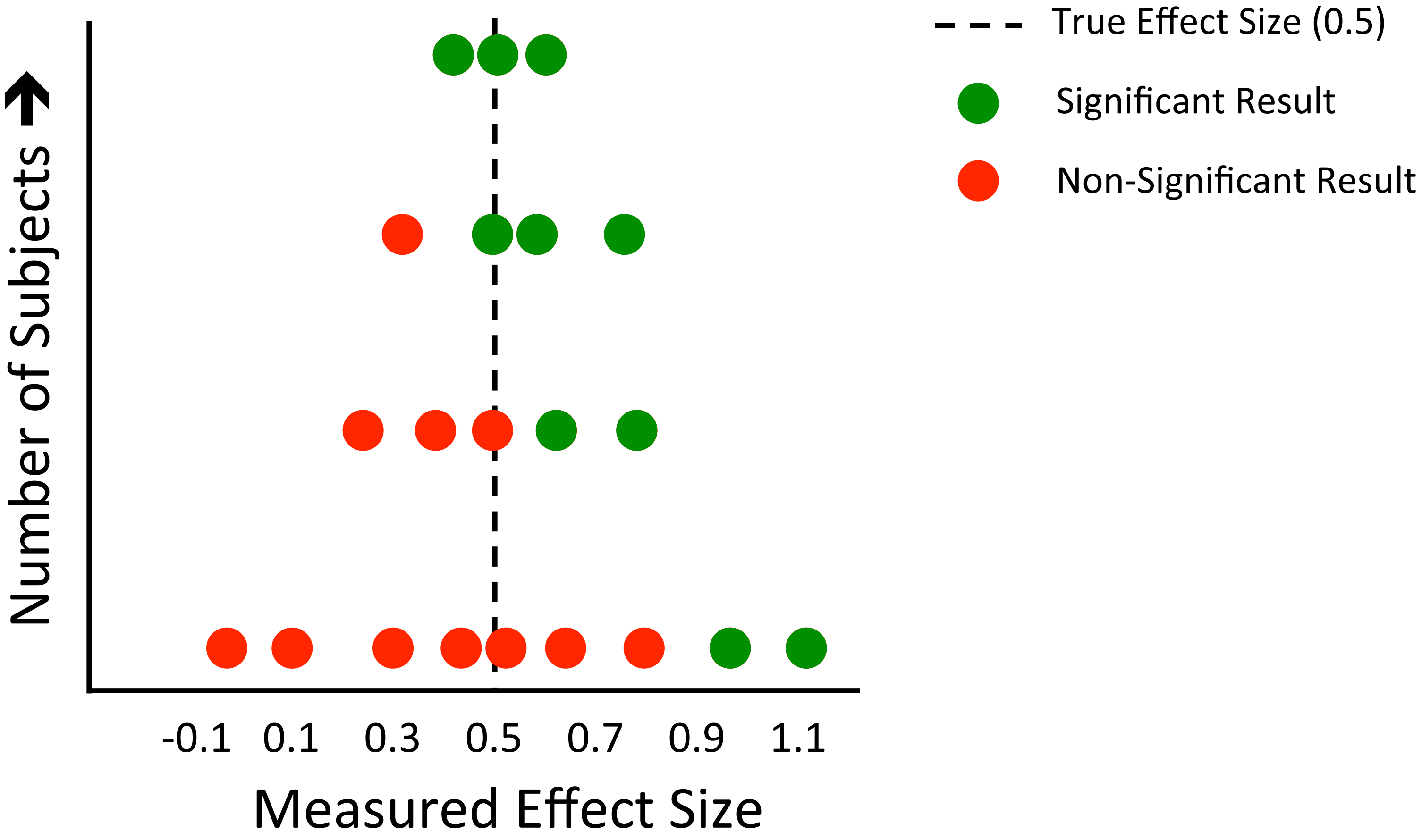
Original study effect size versus replication effect size (correlation coefficients). Diagonal line represents replication effect size equal to original effect size. Dotted line represents replication effect size of 0. Points below the dotted line were effects in the opposite direction of the original. Density plots are separated by significant (blue) and nonsignificant (red) effects.

How do Underpowered Studies Lead to Failed Replications?

- Common misconception: underpowered studies are more likely to result in false positives
  - For a null effect, the false alarm rate is .05 for any sample size (assuming good experimental practices)

- Actual problem: with underpowered studies, only results with really big measured effect sizes will reach significance.
  - This means significant results with small sample sizes are either:
    - A) actually really large effects, or
    - B) small/medium effects that, due to noise, were measured larger than they really are.

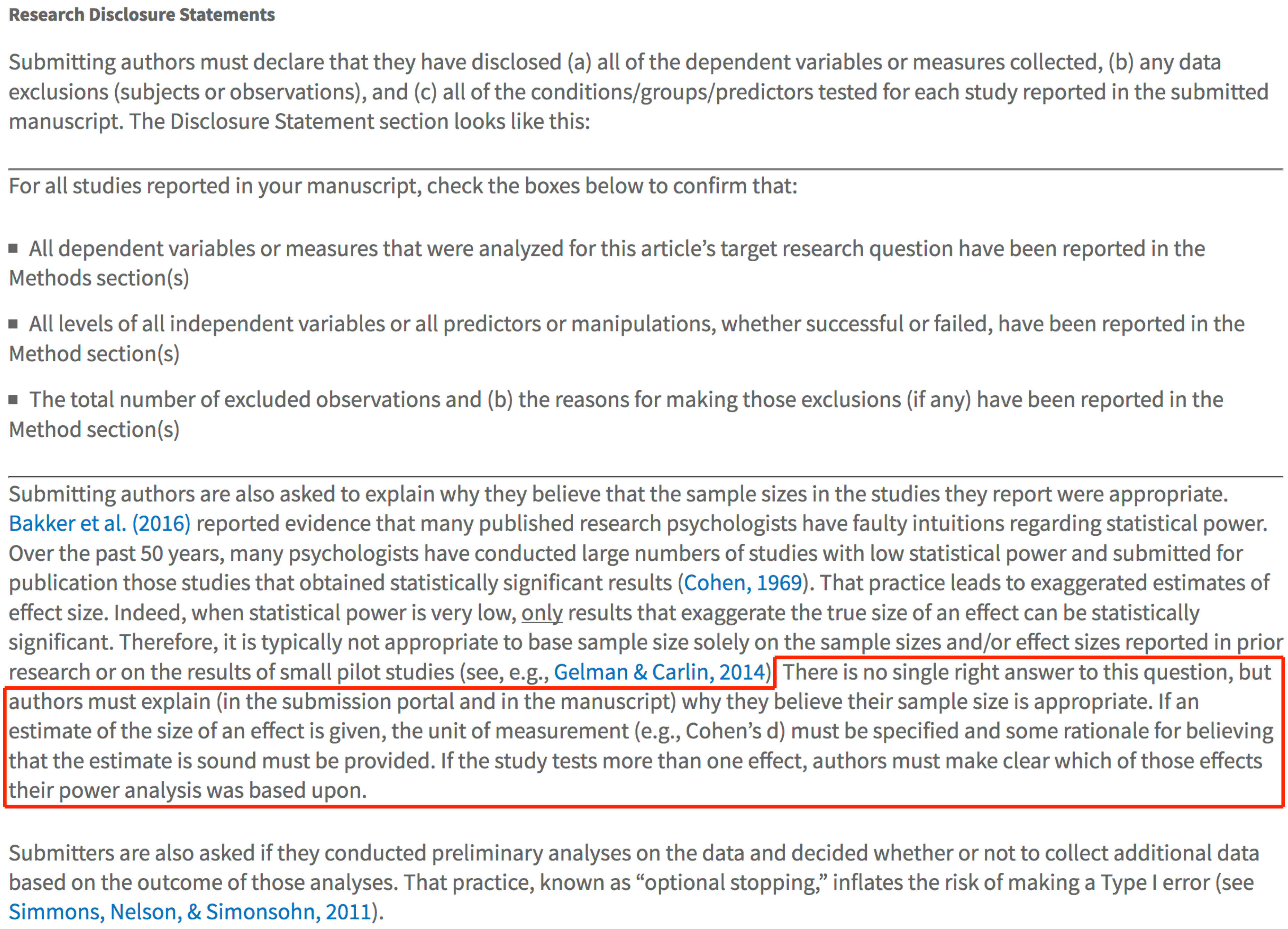
  - When researchers replicate an underpowered study (case B), they will select their sample size based upon the exaggerated effect size, leading to the replication being underpowered (like the original study). Odds are the noise won't lead to an exaggerated effect size again, leading to a failed replication



- Fewer subjects = more variability & lower power.
- Even if your effect is "real," with small N your results will only be significant if you "get lucky" and overestimate the magnitude of the effect (green dots bottom row).
- When others replicate you, they will be underpowered, leading to failed replications (even for real effects)

How to Select Your Sample Size? Guidelines from journals a bit unclear…

How to Select Your Sample Size: Guidelines from *Psych Science*



Research Disclosure Statements

Submitting authors must declare that they have disclosed (a) all of the dependent variables or measures collected, (b) any data exclusions (subjects or observations), and (c) all of the conditions/groups/predictors tested for each study reported in the submitted manuscript. The Disclosure Statement section looks like this:

For all studies reported in your manuscript, check the boxes below to confirm that:

■ All dependent variables or measures that were analyzed for this article's target research question have been reported in the Methods section(s)

■ All levels of all independent variables or all predictors or manipulations, whether successful or failed, have been reported in the Method section(s)

■ The total number of excluded observations and (b) the reasons for making those exclusions (if any) have been reported in the Method section(s)

Submitting authors are also asked to explain why they believe that the sample sizes in the studies they report were appropriate. Bakker et al. (2016) reported evidence that many published research psychologists have faulty intuitions regarding statistical power. Over the past 50 years, many psychologists have conducted large numbers of studies with low statistical power and submitted for publication those studies that obtained statistically significant results (Cohen, 1969). That practice leads to exaggerated estimates of effect size. Indeed, when statistical power is very low, only results that exaggerate the true size of an effect can be statistically significant. Therefore, it is typically not appropriate to base sample size solely on the sample sizes and/or effect sizes reported in prior research or on the results of small pilot studies (see, e.g., Gelman & Carlin, 2014). There is no single right answer to this question, but authors must explain (in the submission portal and in the manuscript) why they believe their sample size is appropriate. If an estimate of the size of an effect is given, the unit of measurement (e.g., Cohen's d) must be specified and some rationale for believing that the estimate is sound must be provided. If the study tests more than one effect, authors must make clear which of those effects their power analysis was based upon.

Submitters are also asked if they conducted preliminary analyses on the data and decided whether or not to collect additional data based on the outcome of those analyses. That practice, known as "optional stopping," inflates the risk of making a Type I error (see Simmons, Nelson, & Simonsohn, 2011).

How to Select Your Sample Size: Guidelines from a *JEP: General* Editor's Commentary

If I want to get published, what are the most important aspects of my research for me to pay attention to?

Reviewers will call you on problems at any level, from the motivation for your work to the error bars on your figures. Cutting corners at any step of the research process will make the paper more difficult to write and the review process more complicated. So, do that power analysis, run that control condition, read that literature your advisor said may be relevant. And write to make an impact, not to get published. The best papers are not trying to meet publication threshold, but are rather aiming to produce the best work in their field.

Are there any red flags that usually lead you to reject an article?

At *JEP: G*, we are paying more attention to the power of experiments and we are increasingly uncomfortable with analyses on very small sample sizes. Publication of low-power studies is simply not good for the field in the long run — these studies inflate the rate of both false negatives and false positives in the literature (Ellis, 2010). Whenever possible, a study's sample size would be justified by a power analysis.

How to justify/select your sample size: some concrete options

How to Justify Your Number of Subjects

- If you've already run your study without doing a power analysis => post-hoc power analysis
  - Calculate power based upon your effect size
  - Calculate power based upon an arbitrary theoretical effect size

  - Not recommended: although this approach is useful for estimating how many subjects would have been needed to detect an effect that failed to reach significance, it won't tell you if your significant effect was exaggerated
  - Doing this alone won't help the replication crises

Warning: Make sure you use the correct effect size

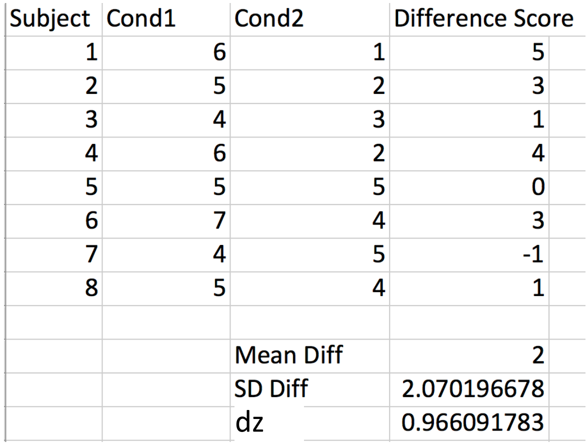
- Between-subjects: Cohen’s d

  $d = (M_1 – M_2) / SD_{pooled}$

$$SD_{pooled} = \sqrt{\frac{\sum(X_1 - \overline{X}_1)^2 + \sum(X_2 - \overline{X}_2)^2}{n_1 + n_2 - 2}} \qquad SD_{pooled} = \sqrt{\frac{(SD_1^2 + SD_2^2)}{2}} \qquad SD^*_{pooled} = \sqrt{\frac{(n_1 - 1)SD_1^2 + (n_2 - 1)SD_2^2}{n_1 + n_2 - 2}}$$

- Within-subjects: Cohen’s dz

  dz = mean(difference_scores)/SD(difference_scores)



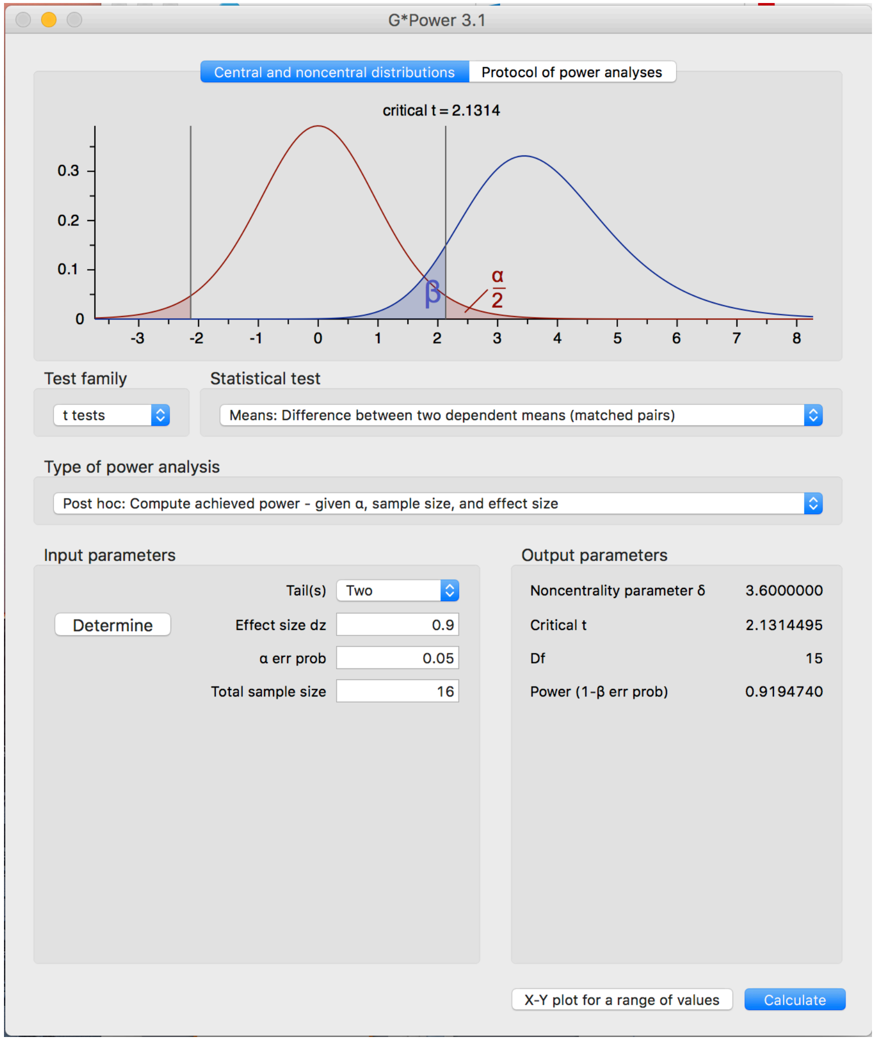
Subject	Cond1	Cond2	Difference Score
1	6	1	5
2	5	2	3
3	4	3	1
4	6	2	4
5	5	5	0
6	7	4	3
7	4	5	-1
8	5	4	1
		Mean Diff	2
		SD Diff	2.070196678
		dz	0.966091783

Post-hoc Power Analysis: G*Power

- Example: I've run a within-subjects study, and found a significant effect with N = 16 and dz = .9

- G*Power tells me that if this is the true effect size, my power was .92

G*Power: gpower.hhu.de



G*Power 3.1

Central and noncentral distributions
Protocol of power analyses
critical t = 2.1314
0.3
0.2
0.1
0
-3
-2
-1
0
1
2
3
4
5
6
7
8
β
α
2

Test family: t tests

Statistical test: Means: Difference between two dependent means (matched pairs)

Type of power analysis: Post hoc: Compute achieved power - given α, sample size, and effect size

Input parameters

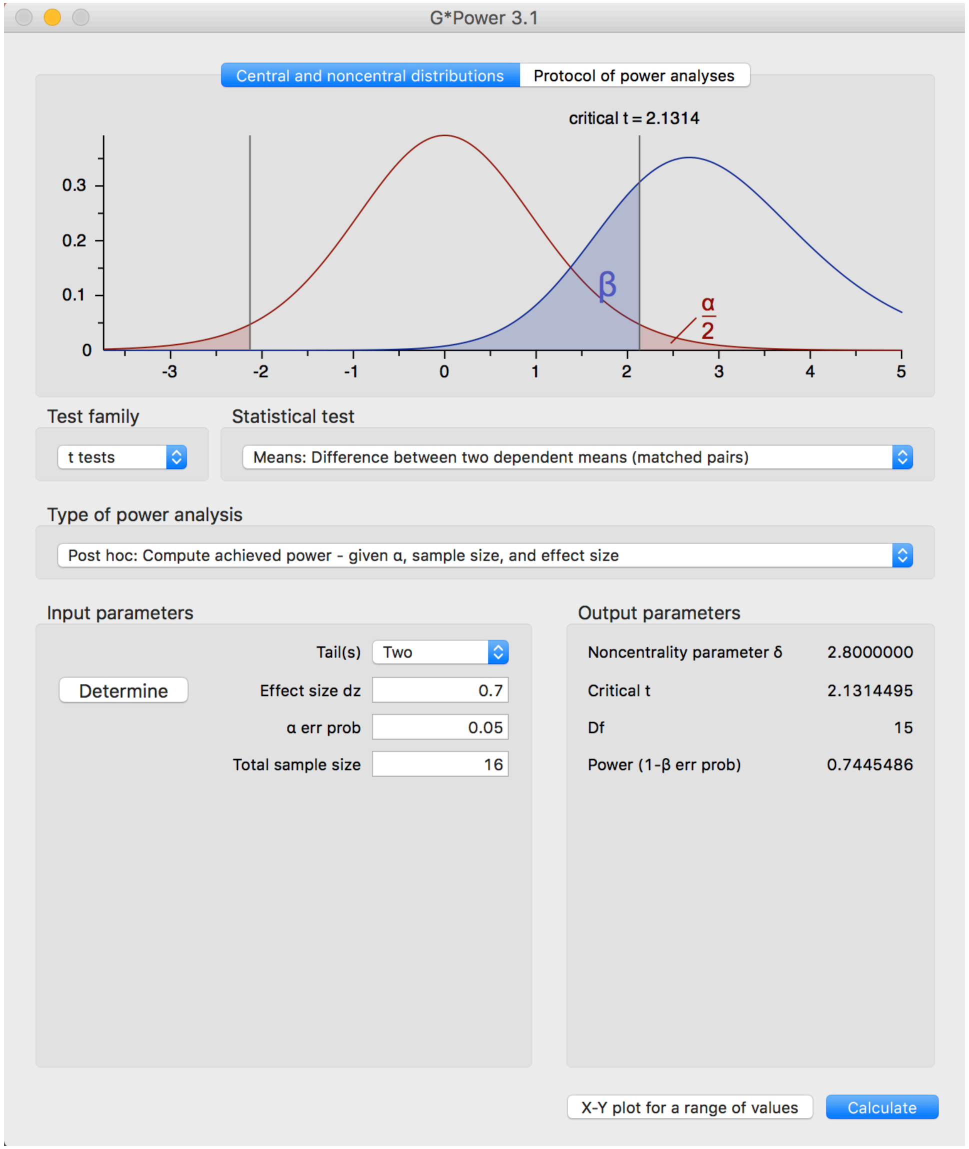
Tail(s)	Two
Effect size dz	0.9
α err prob	0.05
Total sample size	16

Output parameters

Noncentrality parameter δ	3.6000000
Critical t	2.1314495
Df	15
Power (1-β err prob)	0.9194740

Post-hoc Power Analysis

- Example: I've run a within subjects study, found a significant effect with N = 16 and dz = .9
- …but what if my effect was exaggerated? What would my power be at dz = .7?



G*Power 3.1
Central and noncentral distributions
Protocol of power analyses
critical t = 2.1314
0.3
0.2
0.1
0
-3
-2
-1
0
1
2
3
4
5
β
α
2
Determine
X-Y plot for a range of values
Calculate

Test family: t tests

Statistical test: Means: Difference between two dependent means (matched pairs)

Type of power analysis: Post hoc: Compute achieved power – given α, sample size, and effect size

Input parameters

Tail(s): Two

Effect size dz: 0.7

α err prob: 0.05

Total sample size: 16

Output parameters

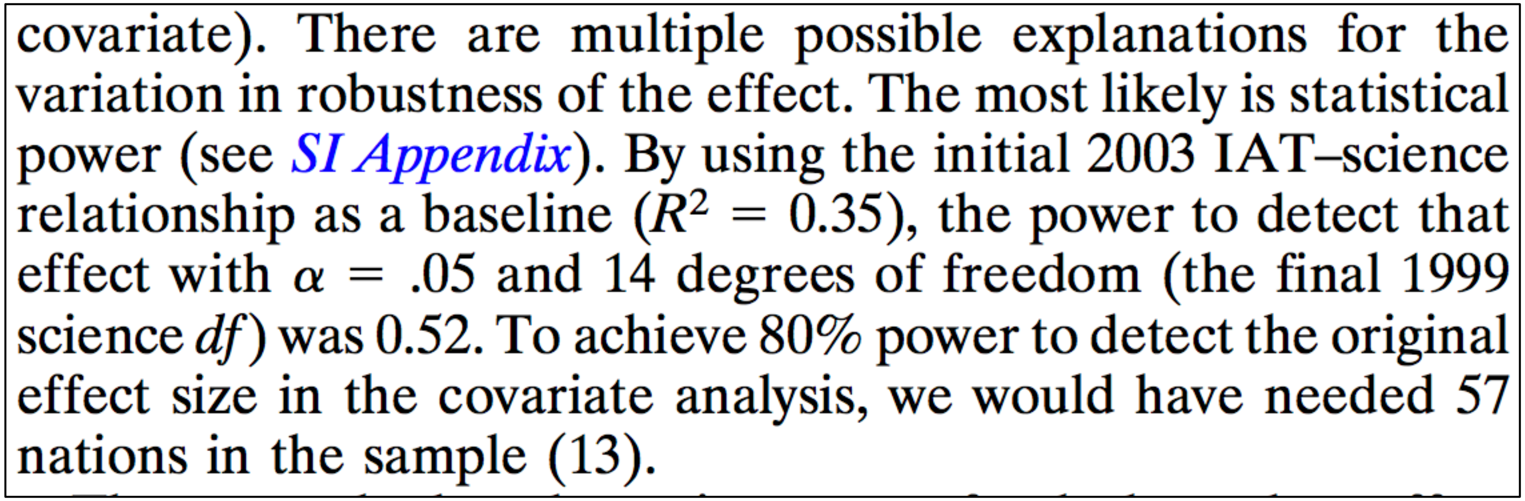
Noncentrality parameter δ: 2.8000000

Critical t: 2.1314495

Df: 15

Power (1-β err prob): 0.7445486

Post-hoc Power Analysis: an example of good usage



covariate). There are multiple possible explanations for the variation in robustness of the effect. The most likely is statistical power (see *SI Appendix*). By using the initial 2003 IAT–science relationship as a baseline ($R^2 = 0.35$), the power to detect that effect with $\alpha = .05$ and 14 degrees of freedom (the final 1999 science *df*) was 0.52. To achieve 80% power to detect the original effect size in the covariate analysis, we would have needed 57 nations in the sample (13).

Nosek et al., 2009. *PNAS*

Better Question:
How to Select Your Number of Subjects?

- If you haven't already collected your data

  => A priori power analysis

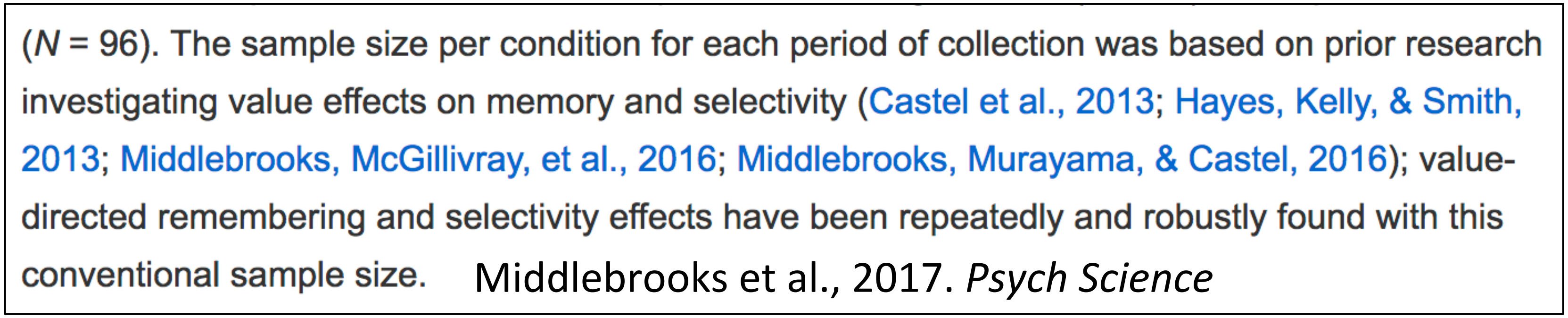
4 Ways to Do This:

1. Use sample size of previous study
2. Calculate number of subjects to run based upon theoretical effect size
3. Calculate number of subjects to run based upon effect size from pilot data
4. Calculate number of subjects to run based upon pilot data & simulations

How to Select Your Number of Subjects?
1. Sample Size of Previous Study

Use sample sizes that have been successful at detecting similar effects in the past

- Not ideal (even small changes between studies, such number of trials, can make a big difference)
- Better than nothing if collecting pilot data isn't feasible



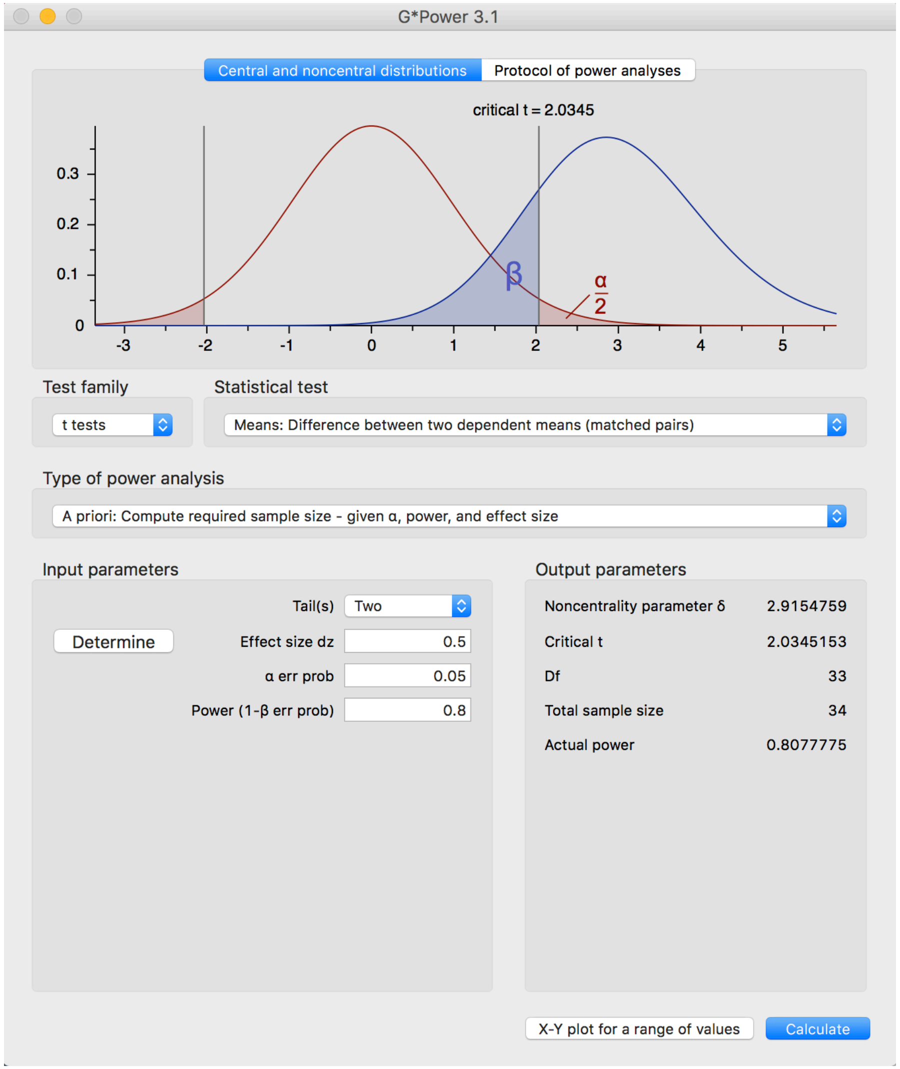
(*N* = 96). The sample size per condition for each period of collection was based on prior research investigating value effects on memory and selectivity (Castel et al., 2013; Hayes, Kelly, & Smith, 2013; Middlebrooks, McGillivray, et al., 2016; Middlebrooks, Murayama, & Castel, 2016); value-directed remembering and selectivity effects have been repeatedly and robustly found with this conventional sample size.

Middlebrooks et al., 2017. *Psych Science*

How to Select Your Number of Subjects?

2. Use a Theoretical Effect Size

- Example: I don't know what the effect size will be, but I want to be able to detect a medium effect size (>= .5) with power = .8.
- Can also use effect sizes from meta-analyses
- Still not ideal (as your study may differ from the average study in meta-analysis), but better


G*Power 3.1
Central and noncentral distributions
Protocol of power analyses
critical t = 2.0345
0.3
0.2
0.1
0
-3
-2
-1
0
1
2
3
4
5
β
α
2
Determine
X-Y plot for a range of values
Calculate

Test family: t tests

Statistical test: Means: Difference between two dependent means (matched pairs)

Type of power analysis: A priori: Compute required sample size - given α, power, and effect size

Input parameters

Tail(s)	Two
Effect size dz	0.5
α err prob	0.05
Power (1-β err prob)	0.8

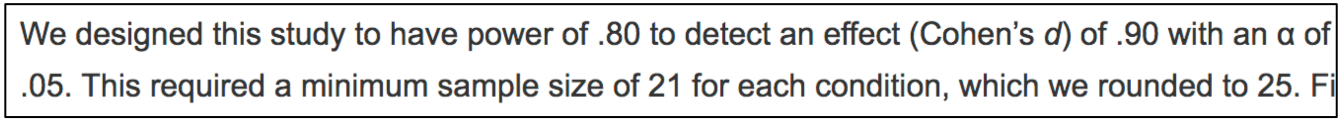
Output parameters

Noncentrality parameter δ	2.9154759
Critical t	2.0345153
Df	33
Total sample size	34
Actual power	0.8077775

How to Select Your Number of Subjects?

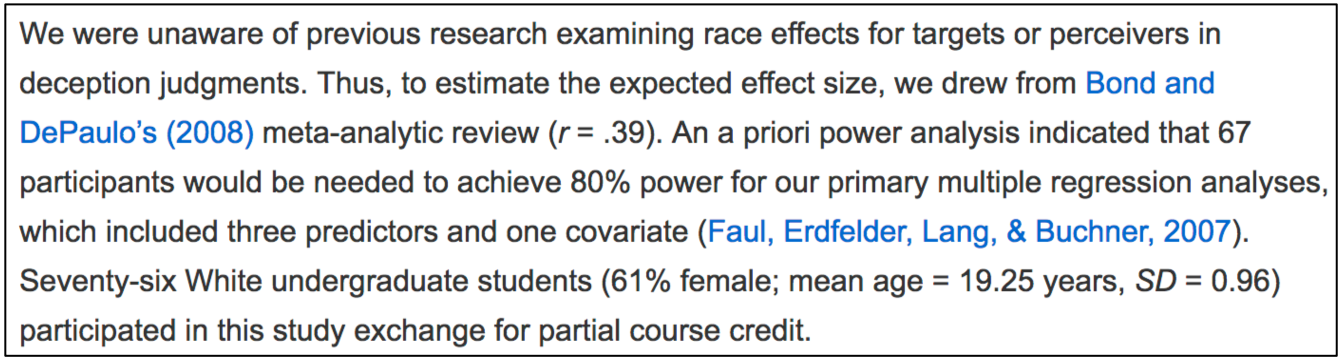
2. Use a Theoretical Effect Size

Park et al., 2017. *Psych Science*

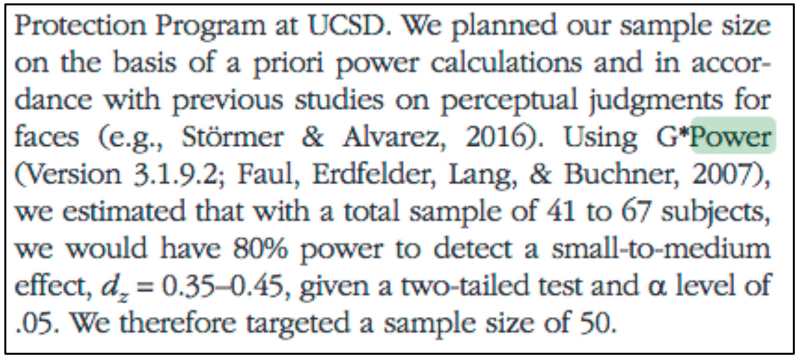


We designed this study to have power of .80 to detect an effect (Cohen's *d*) of .90 with an α of .05. This required a minimum sample size of 21 for each condition, which we rounded to 25. Fi

Lloyd et al., 2017. *Psych Science*



We were unaware of previous research examining race effects for targets or perceivers in deception judgments. Thus, to estimate the expected effect size, we drew from Bond and DePaulo's (2008) meta-analytic review (r = .39). An a priori power analysis indicated that 67 participants would be needed to achieve 80% power for our primary multiple regression analyses, which included three predictors and one covariate (Faul, Erdfelder, Lang, & Buchner, 2007). Seventy-six White undergraduate students (61% female; mean age = 19.25 years, *SD* = 0.96) participated in this study exchange for partial course credit.

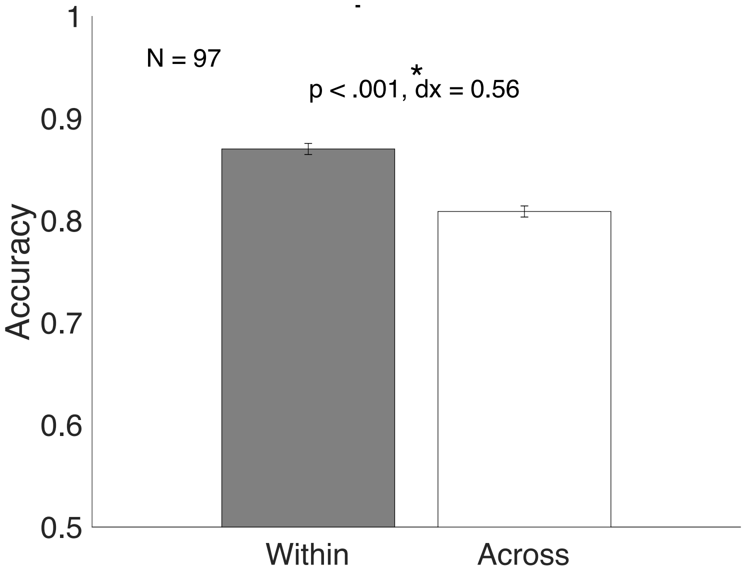


Protection Program at UCSD. We planned our sample size on the basis of a priori power calculations and in accordance with previous studies on perceptual judgments for faces (e.g., Störmer & Alvarez, 2016). Using G*Power (Version 3.1.9.2; Faul, Erdfelder, Lang, & Buchner, 2007), we estimated that with a total sample of 41 to 67 subjects, we would have 80% power to detect a small-to-medium effect, d_z = 0.35–0.45, given a two-tailed test and α level of .05. We therefore targeted a sample size of 50.

Carr et al., 2017. *Psych Science*

How to Select Your Number of Subjects?

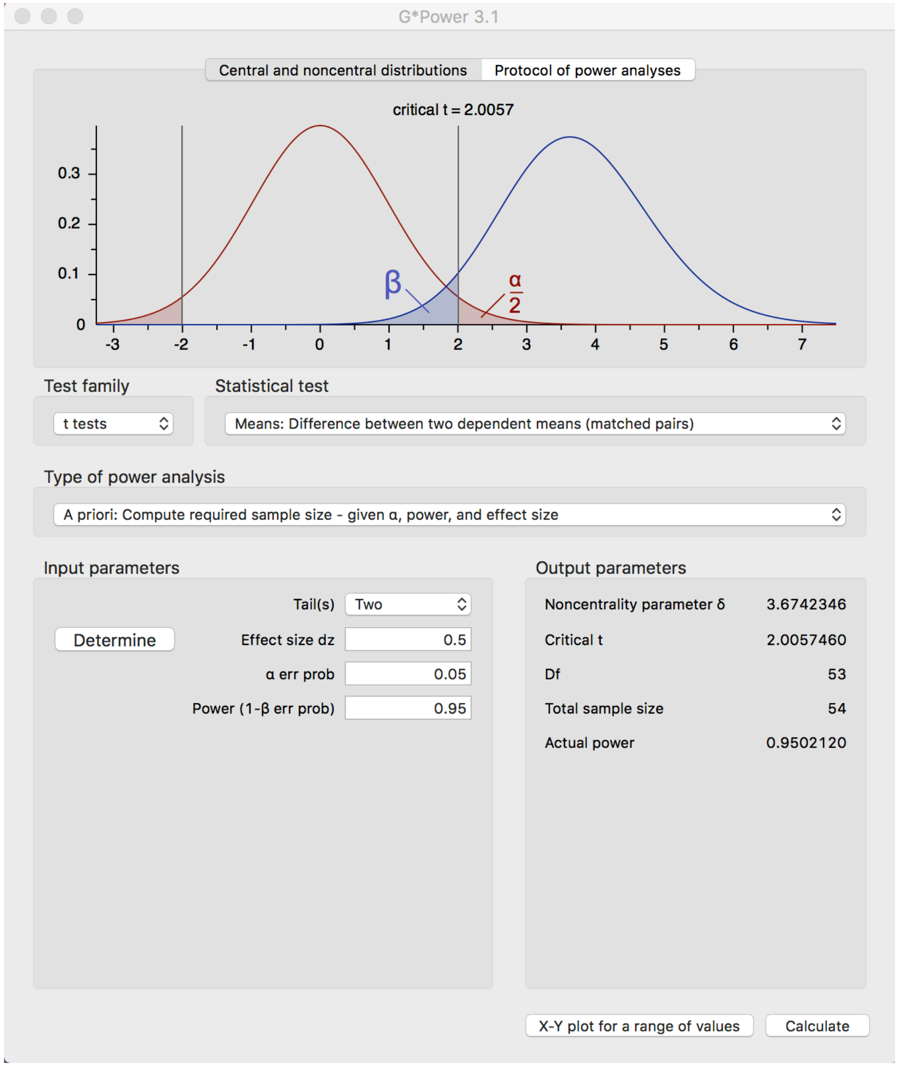
3. Use Effect Size From Pilot Data



- If you can easily get data, run a big pilot set! This is pilot data from a 10 minute online study (don't need this many subjects in pilot, but more the better for estimating effect size accurately)

- Pilot data effect size was .56…I told G*power I wanted 95% power to detect .5 effect size, which requires 54 participants for main data set. I rounded up to 60 subjects for the main data set.

- Even this approach isn't perfect – G*power does not take number of trials into account – unless you have hundreds of trials per condition your actual power will be lower than this


G*Power 3.1
Central and noncentral distributions
Protocol of power analyses
critical t = 2.0057
0.3
0.2
0.1
0
-3
-2
-1
0
1
2
3
4
5
6
7
β
α/2
Determine
X-Y plot for a range of values
Calculate

Test family: t tests

Statistical test: Means: Difference between two dependent means (matched pairs)

Type of power analysis: A priori: Compute required sample size - given α, power, and effect size

Input parameters

Tail(s)	Two
Effect size dz	0.5
α err prob	0.05
Power (1-β err prob)	0.95

Output parameters

Noncentrality parameter δ	3.6742346
Critical t	2.0057460
Df	53
Total sample size	54
Actual power	0.9502120

How to Select Your Number of Subjects?
4. Pilot Data + Simulation

- Although previous examples give you a decent estimate of power, simulation allows you to calculate power for your exact design (number of trials, comparisons you are interested in, etc.)

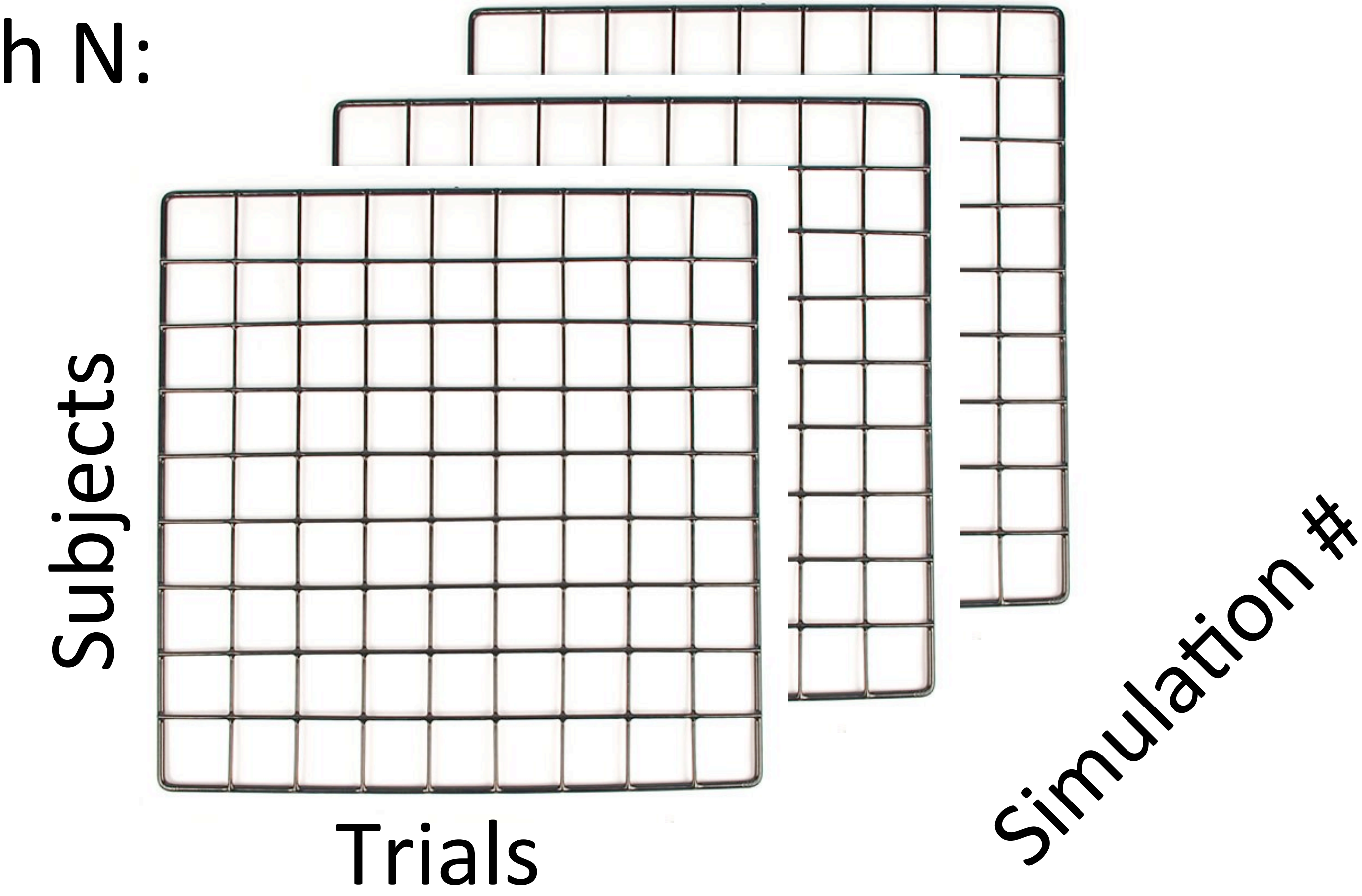
- Simulations can also be used to calculate power for different number of trials than your pilot data

How to Select Your Number of Subjects?

4. Pilot Data + Simulation

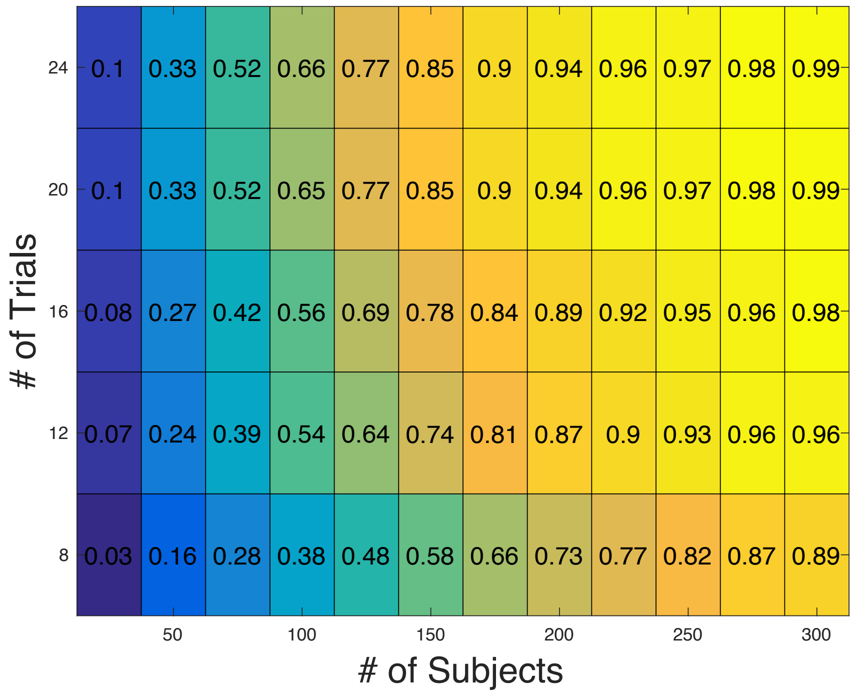
- For each simulated N, sample randomly with replacement from your subjects. Also, sample trials randomly (with replacement) for each subject.

For each N:



How to Select Your Number of Subjects?

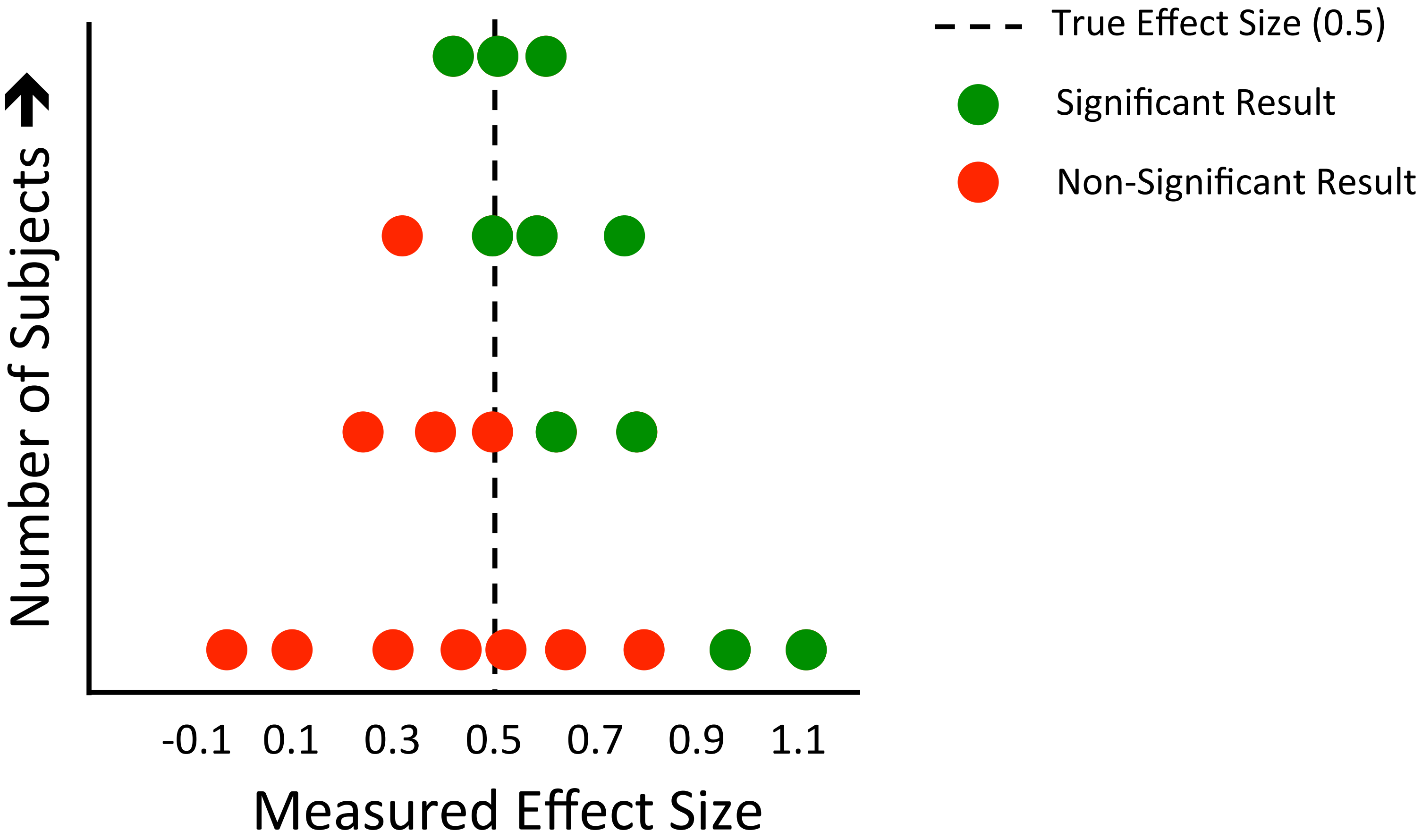
4. Pilot Data + Simulation



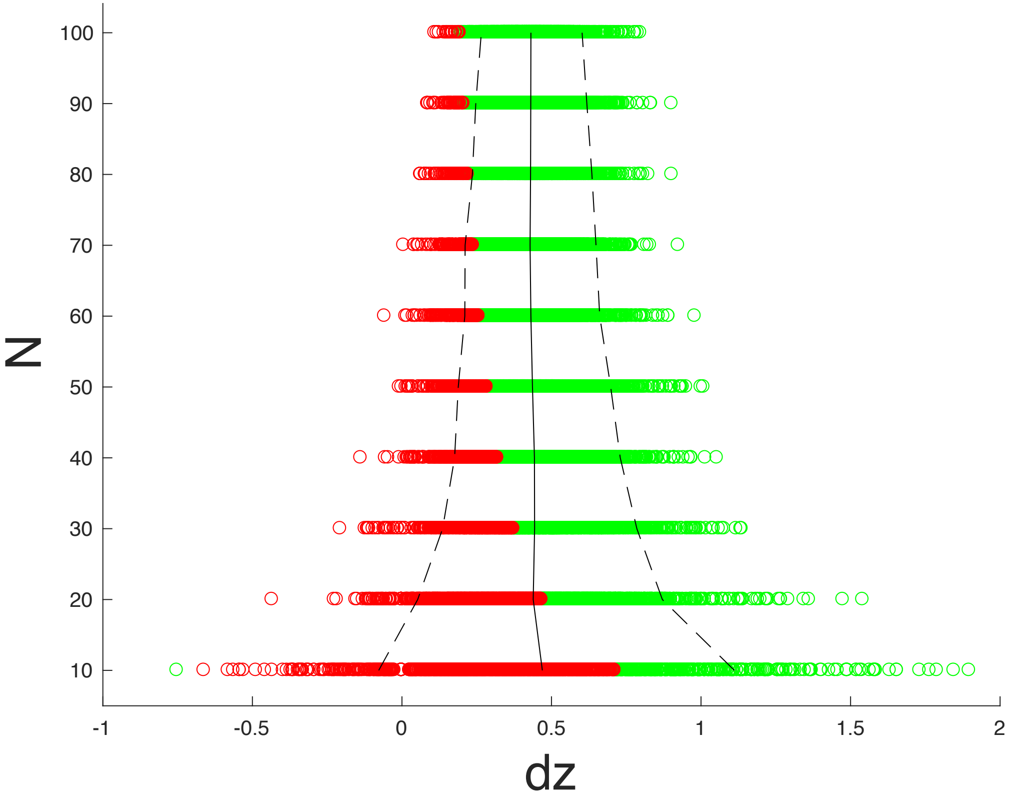
# of Trials		50		100		150		200		250		300
24	0.1	0.33	0.52	0.66	0.77	0.85	0.9	0.94	0.96	0.97	0.98	0.99
20	0.1	0.33	0.52	0.65	0.77	0.85	0.9	0.94	0.96	0.97	0.98	0.99
16	0.08	0.27	0.42	0.56	0.69	0.78	0.84	0.89	0.92	0.95	0.96	0.98
12	0.07	0.24	0.39	0.54	0.64	0.74	0.81	0.87	0.9	0.93	0.96	0.96
8	0.03	0.16	0.28	0.38	0.48	0.58	0.66	0.73	0.77	0.82	0.87	0.89

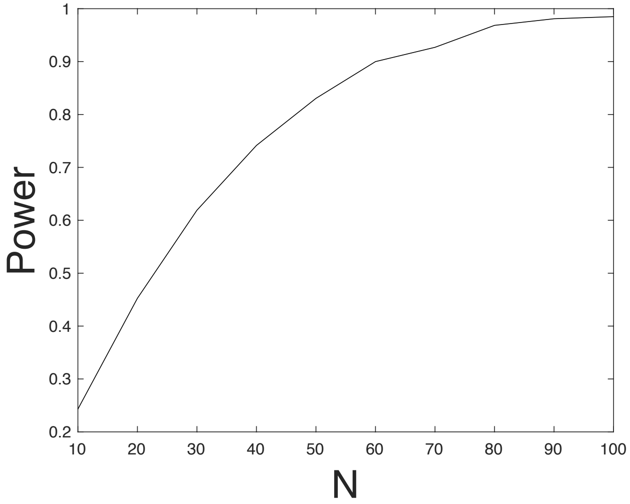
- Using a pilot set of 30 subjects who did 16 trials per condition, simulation allowed me to calculate power for various combinations of trials and subjects. I ended up running the full study using 270 subjects with 12 trials per condition (approximately .95 power)

Do we get this pattern when we simulate using real data?



Simulation Results





Notes:

- Resampling from pilot data
- Power increases with increasing N
- Variability decreases with increasing N
- Average effect size does not change much with N
- Only large, exaggerated effect sizes reach significance at small N
  - Even get one significant result in wrong direction at N = 10 (green dot far left)

Summary

- Small sample sizes are bad for two reasons:
    1. Low power to detect real effects
    2. Any significant result may be exaggerated, leading to failed replications

- When possible, collect a moderately sized pilot set of data to guide your sample size