Power analysis

Psych 251

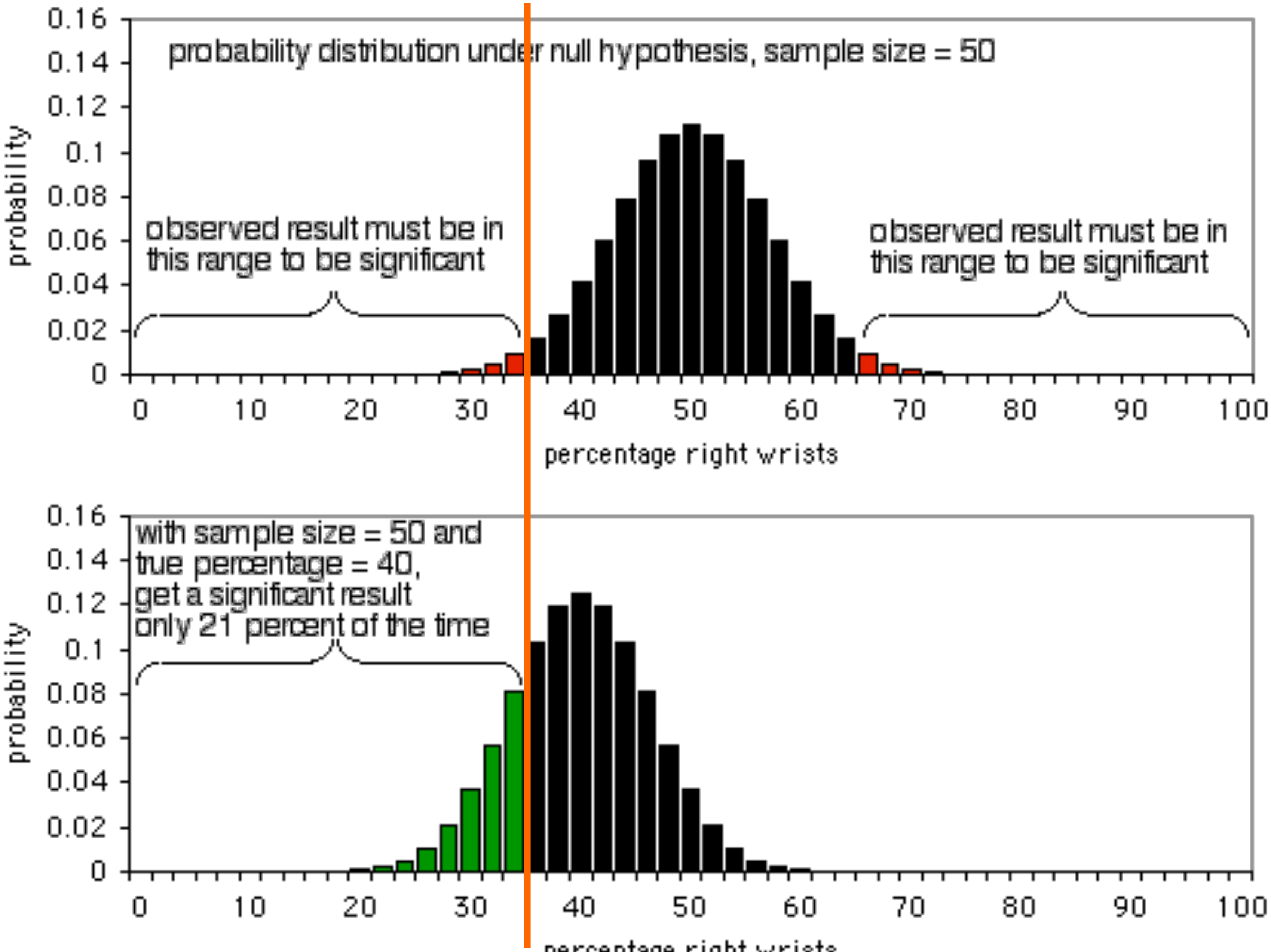
10/25/2017

Classic NHST power

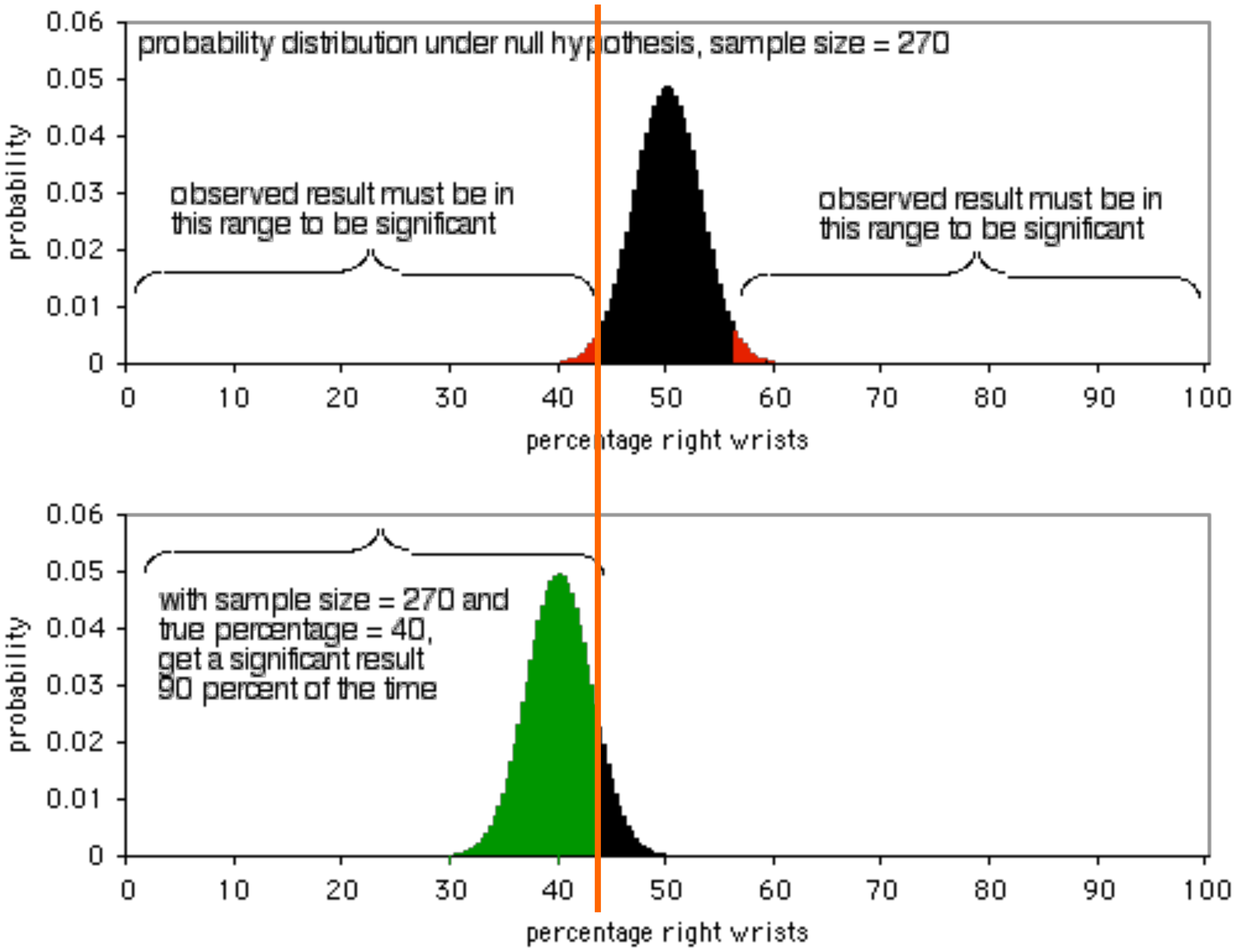
		Truth	
		Null is true	Null is false
Experimental result	Null is false	**α** **Type I error** False positive	Correct
	Null is true	Correct	**β** **Type II error** False negative

Power is 1- β

Quick illustration



Quick illustration



Probing power intuitions

- T-test
  - D=.5 (average in psychology), N_{total}=24
  - D=.8, N=36
  - D=.3, N=120
- Correlation
  - r = .5, N=36
  - r = .3, N=80

A Priori Power Analysis

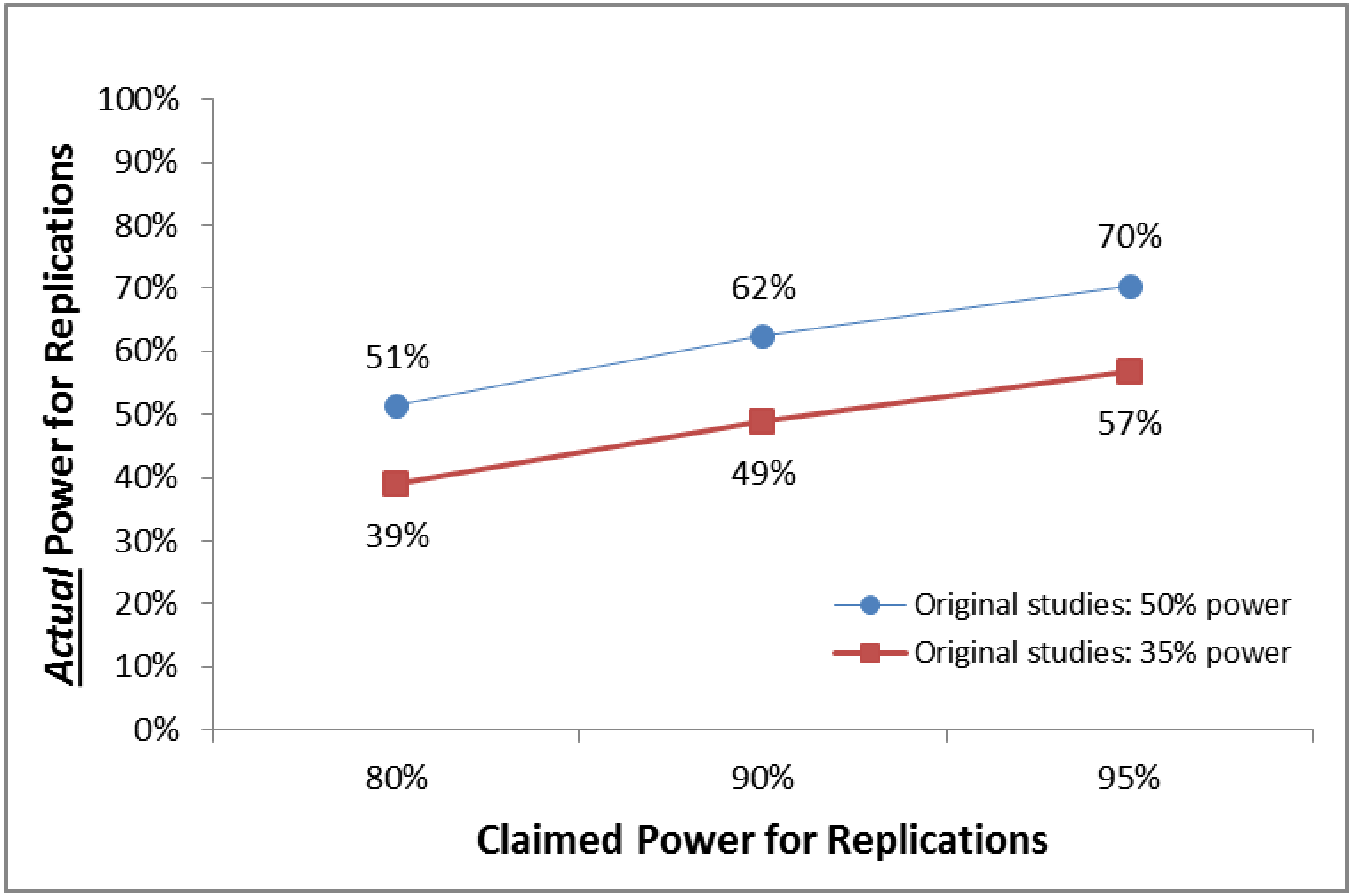
- You want to find how many cases you will need to have a specified amount of power given
  - a specified effect size
  - the criterion of significance to be employed
  - whether the hypotheses are directional or non-directional
- Important part of the planning of research
  - Especially large-scale measurement/intervention studies
  - When arguing for feasibility

A Posteriori Power Analysis

- In principle: Just moving the algebra around
- You want to find out what power would be for a specified
  - effect size
  - sample size
  - and type of analysis
- But power analysis is best done as part of the planning of research
  - could be done after the research to tell you what you should have known earlier
- **Real problems with post-hoc power!**

Post-hoc power suffers from bias

Figure 3. Actual vs Claimed Power in Replications When Sample Size is Based on Observed Effect Size

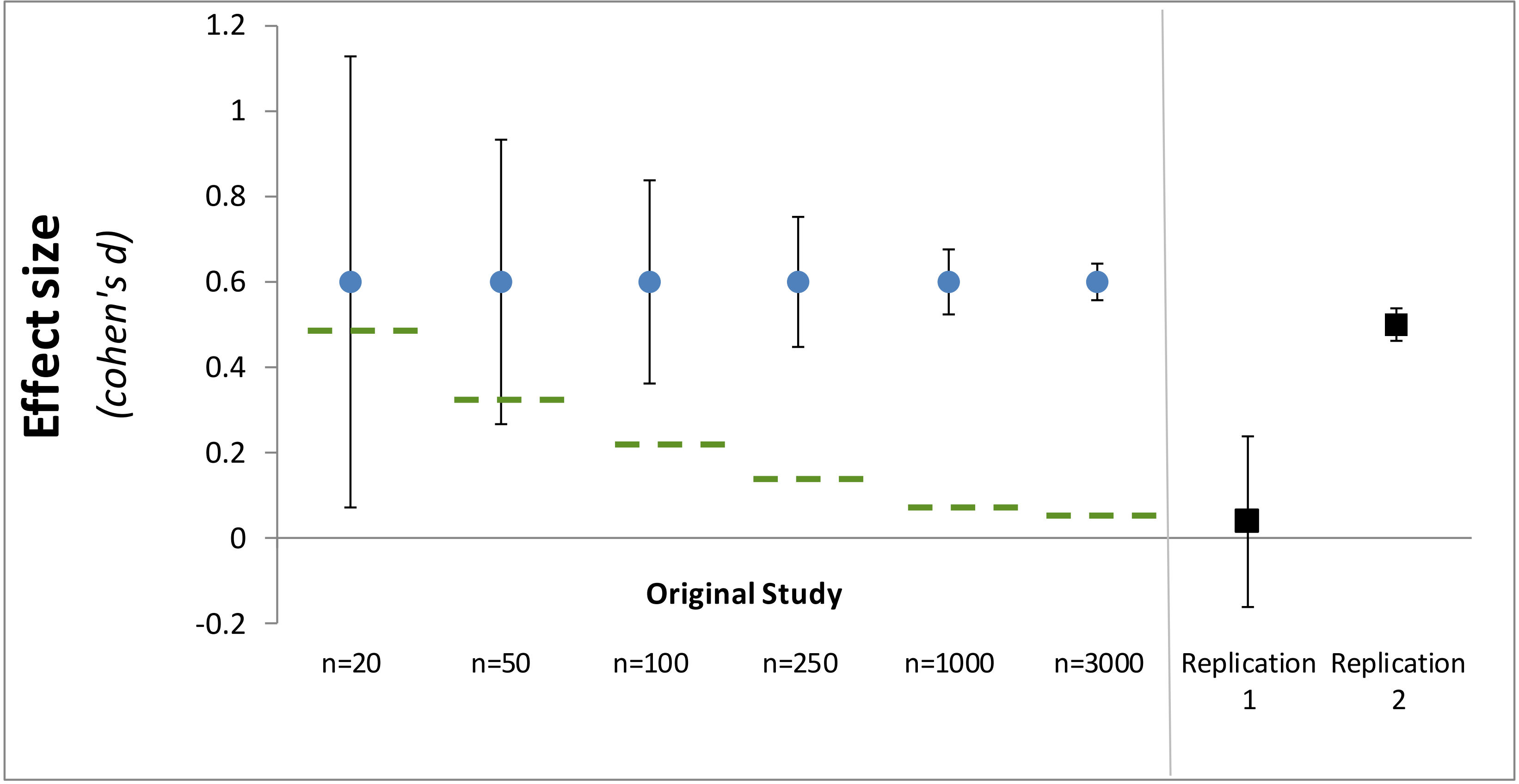


Simonsohn (in press)

“Small telescopes” approach (Simonsohn)

If the original study couldn’t see an effect with power > 33%, it’s probably not the same effect.

Figure 5. Sample size of original study, confidence intervals, and smallest detectable effect ($d_{33\%}$)



The chart depicts results from hypothetical studies, all originals obtaining $\hat{d}$=0.6 with increasingly larger sample sizes. Vertical lines correspond to 95% confidence intervals.

Small telescopes standard

- “a replication needs 2.5 times as many observations as the original study to have about 80% power to reject $d_{33\%}$. For example, if the true effect is zero, and an original study had 20 observations per cell, a replication with 50 observations per cell would have about 80% power to reject the hypothesis that the effect is $d_{33\%}$”

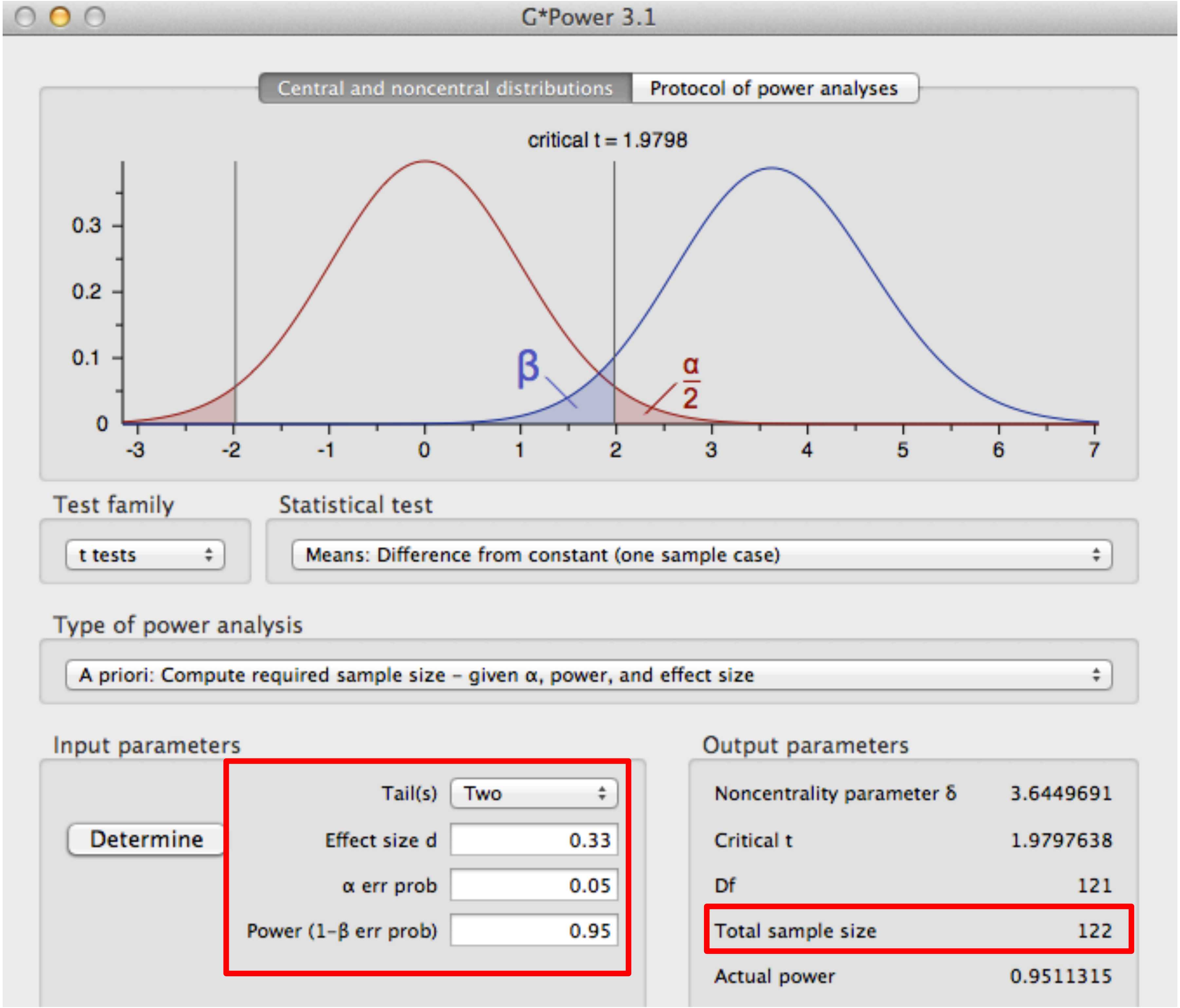
Simonsohn, 2015

Summary: Power analysis

- Want to find out how many subjects to run
- Need three things
  - Alpha value (usually .05)
  - Required power (often .8, .9, or .95)
  - Effect size (this is the hard one)
- How do we get effect size
  - Post-hoc: just read it off (well…)
  - A priori: estimate from the literature, check a range, run a pilot study

G*Power

- Software for computing effect sizes
- Statistical test
  - Many options here: f, t, chi2, r, binomial
- Type of power analysis
  - A priori
  - Post-hoc
  - Others
- Input parameters and press go
- Like all statistical software, easy to get the wrong answer!



Test family

t tests

Statistical test

Means: Difference from constant (one sample case)

Type of power analysis

A priori: Compute required sample size – given α, power, and effect size

Input parameters

Tail(s) Two

Effect size d 0.33

α err prob 0.05

Power (1–β err prob) 0.95

Output parameters

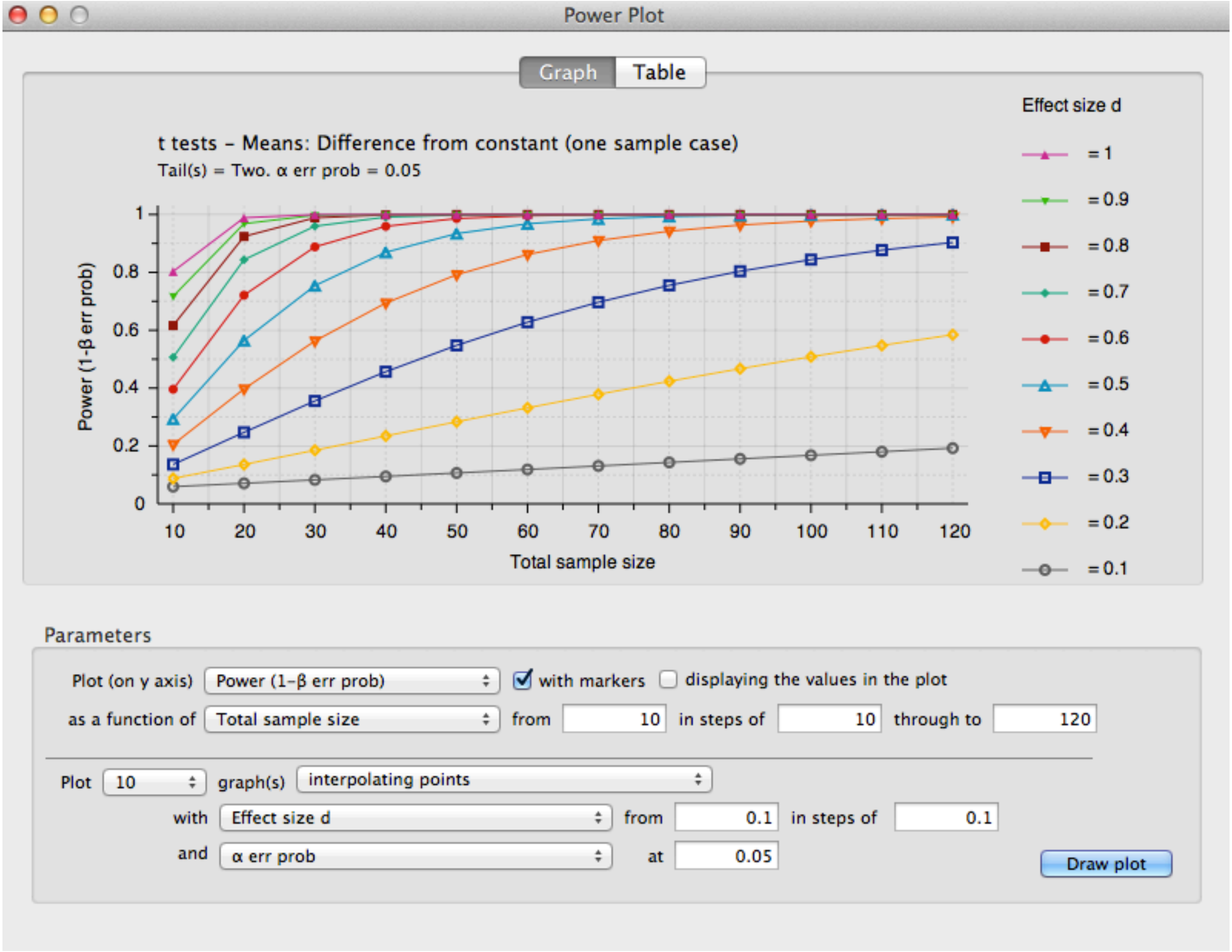
Noncentrality parameter δ 3.6449691

Critical t 1.9797638

Df 121

Total sample size 122

Actual power 0.9511315



Parameters

Plot (on y axis) Power (1–β err prob) ☑ with markers □ displaying the values in the plot

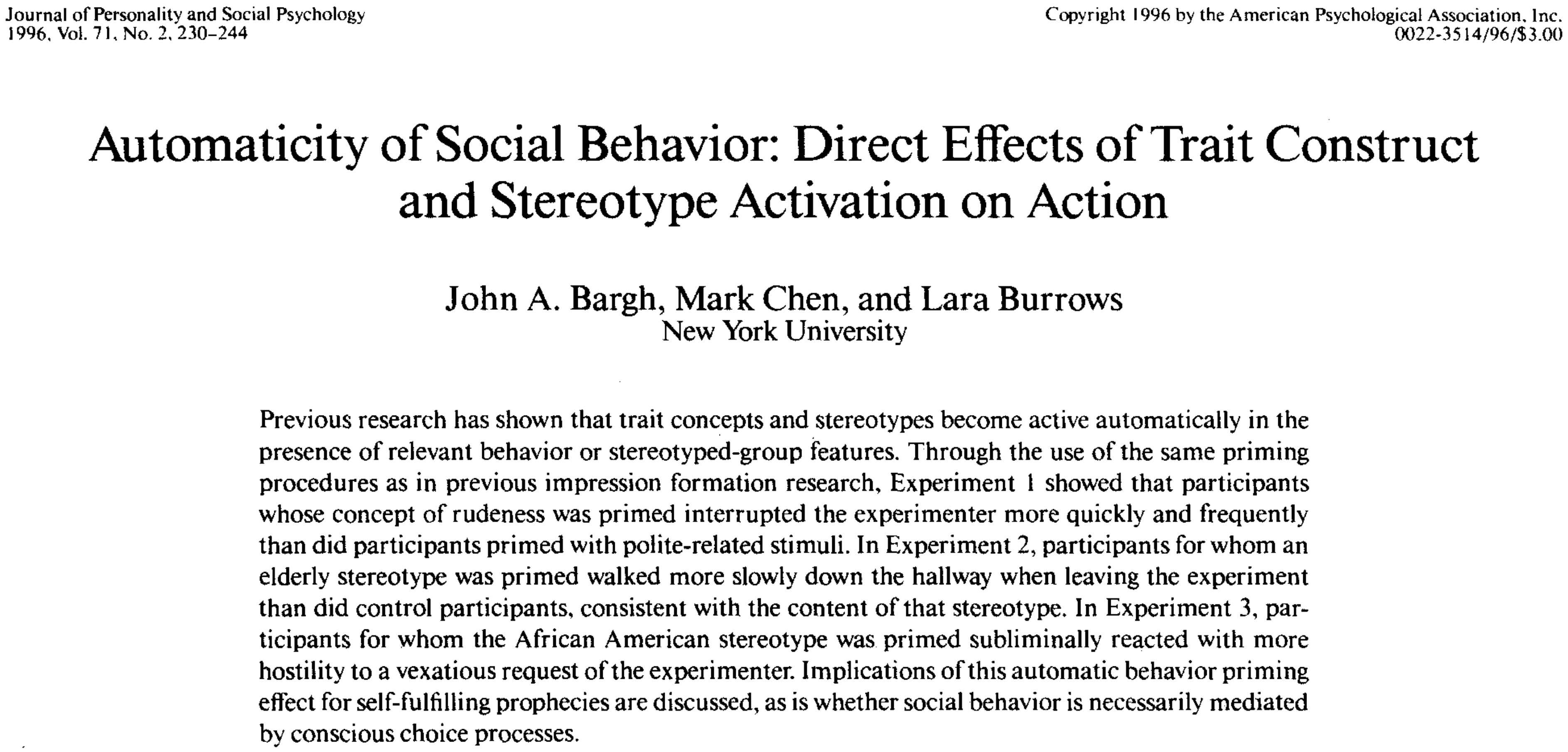
as a function of Total sample size from 10 in steps of 10 through to 120

Plot 10 graph(s) interpolating points

with Effect size d from 0.1 in steps of 0.1

and α err prob at 0.05

Slightly more complex



Journal of Personality and Social Psychology
1996, Vol. 71, No. 2, 230–244

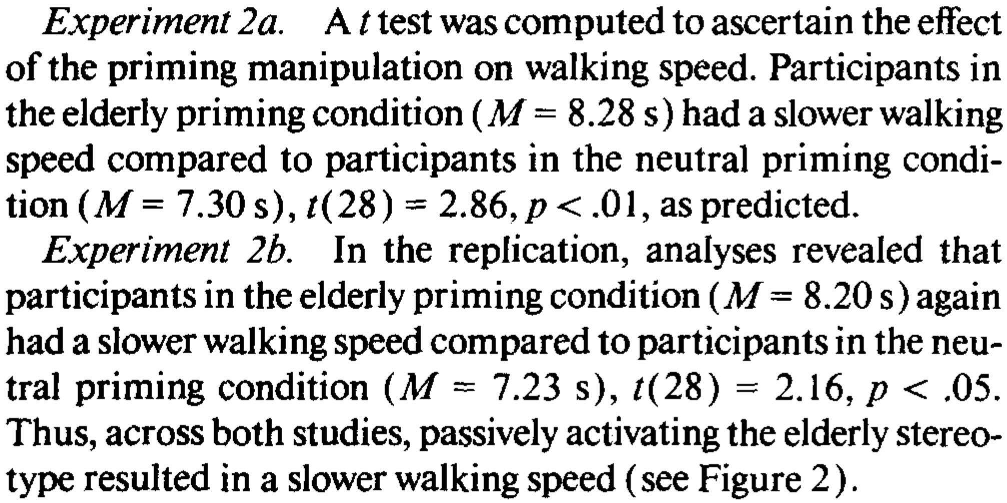
Automaticity of Social Behavior: Direct Effects of Trait Construct and Stereotype Activation on Action

John A. Bargh, Mark Chen, and Lara Burrows
New York University

Previous research has shown that trait concepts and stereotypes become active automatically in the presence of relevant behavior or stereotyped-group features. Through the use of the same priming procedures as in previous impression formation research, Experiment 1 showed that participants whose concept of rudeness was primed interrupted the experimenter more quickly and frequently than did participants primed with polite-related stimuli. In Experiment 2, participants for whom an elderly stereotype was primed walked more slowly down the hallway when leaving the experiment than did control participants, consistent with the content of that stereotype. In Experiment 3, participants for whom the African American stereotype was primed subliminally reacted with more hostility to a vexatious request of the experimenter. Implications of this automatic behavior priming effect for self-fulfilling prophecies are discussed, as is whether social behavior is necessarily mediated by conscious choice processes.

Slightly more complex

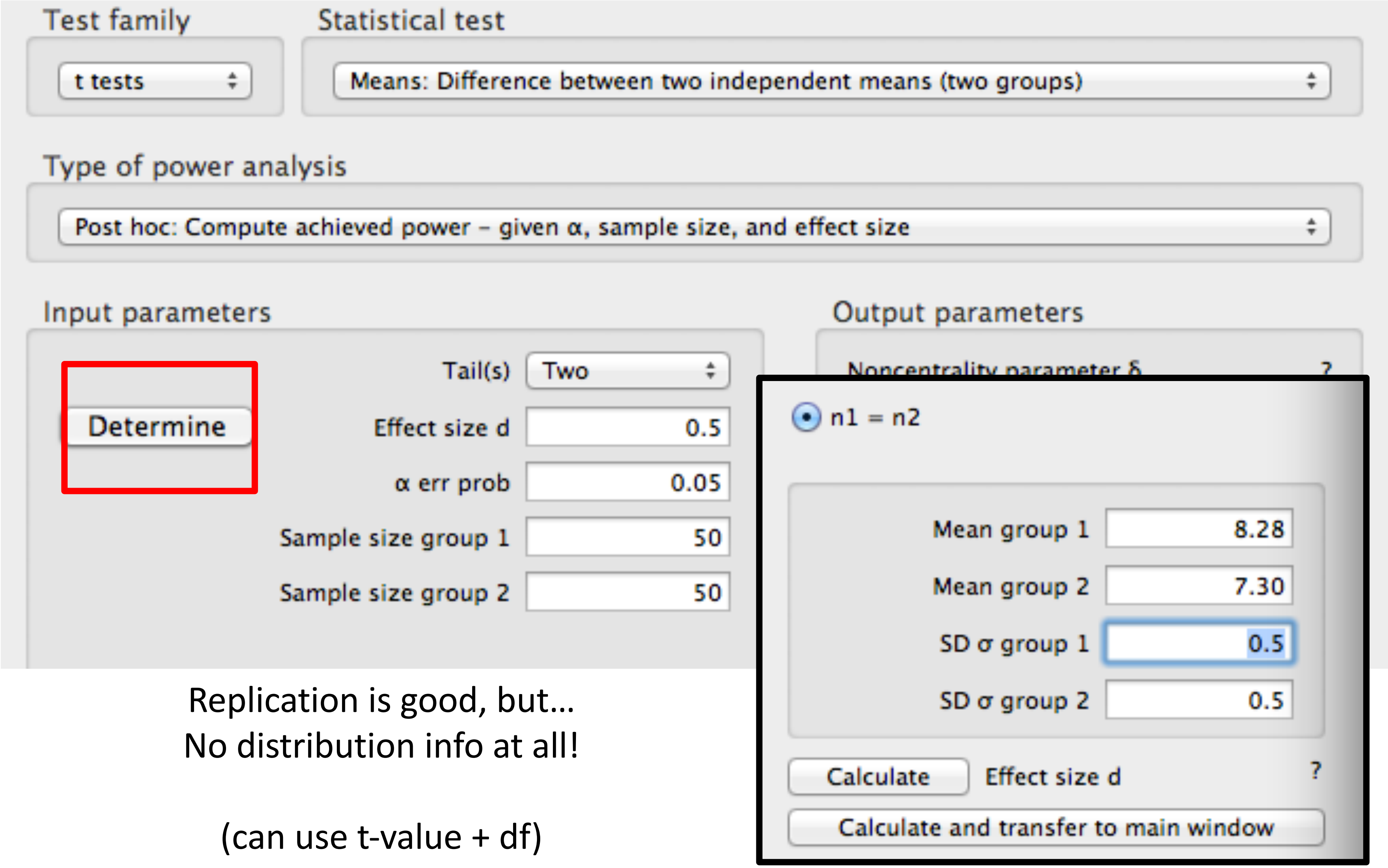
- Bargh, Chen, & Burrows (1996)
- Primed participants by having them unscramble sentences about being elderly
  - Worried, Florida, old, lonely …



Experiment 2a. A t test was computed to ascertain the effect of the priming manipulation on walking speed. Participants in the elderly priming condition ($M = 8.28$ s) had a slower walking speed compared to participants in the neutral priming condition ($M = 7.30$ s), $t(28) = 2.86, p < .01$, as predicted.

Experiment 2b. In the replication, analyses revealed that participants in the elderly priming condition ($M = 8.20$ s) again had a slower walking speed compared to participants in the neutral priming condition ($M = 7.23$ s), $t(28) = 2.16, p < .05$. Thus, across both studies, passively activating the elderly stereotype resulted in a slower walking speed (see Figure 2).

Slightly more complex



Test family: t tests

Statistical test: Means: Difference between two independent means (two groups)

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

Input parameters

Tail(s)	Two
Effect size d	0.5
α err prob	0.05
Sample size group 1	50
Sample size group 2	50

Output parameters

Noncentrality parameter δ ?

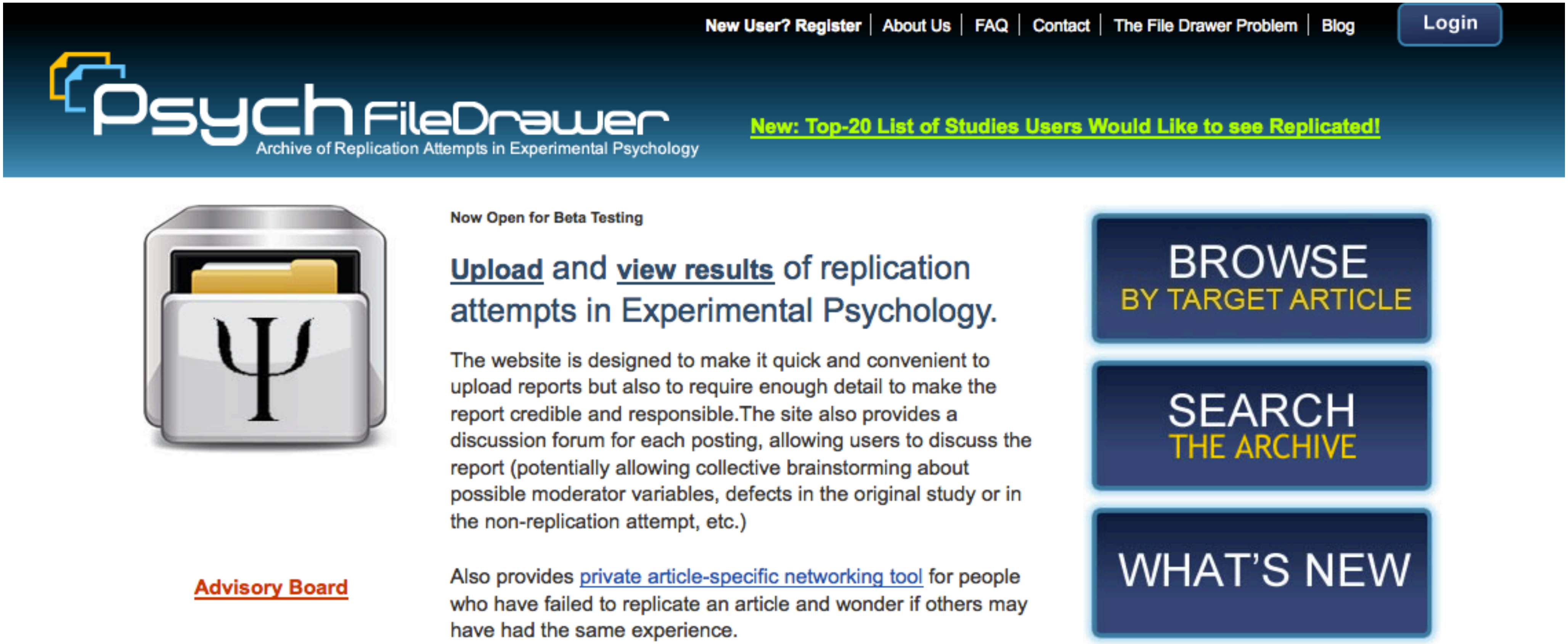
● n1 = n2

Mean group 1	8.28
Mean group 2	7.30
SD σ group 1	0.5
SD σ group 2	0.5

Effect size d ?

Replication is good, but…
No distribution info at all!

(can use t-value + df)



Now Open for Beta Testing

Upload and view results of replication attempts in Experimental Psychology.

The website is designed to make it quick and convenient to upload reports but also to require enough detail to make the report credible and responsible.The site also provides a discussion forum for each posting, allowing users to discuss the report (potentially allowing collective brainstorming about possible moderator variables, defects in the original study or in the non-replication attempt, etc.)

Also provides private article-specific networking tool for people who have failed to replicate an article and wonder if others may have had the same experience.

A t-test revealed that participants in the elderly priming condition walked faster (M=7.82, SD=1.03) than subjects in the neutral priming condition (M=8.06, SD=1.15). This difference was not significant, t(64) = -.88, p=.38.

(replication by Pashler et al.)

Computed power

- Based on t→ d, ES is *d*=1.08
- Based on Pashler et al. SDs, effect size in original would also have been around 1, with around 1s difference between means and 1s SD from replication
- So .95 power at alpha = .05 would be around 68 subjects
- Pashler et al. replication had n=66…

OPEN ACCESS Freely available online

PLoS one

Behavioral Priming: It's all in the Mind, but Whose Mind?

Stéphane Doyen[1,2,3]*, Olivier Klein[2], Cora-Lise Pichon[1], Axel Cleeremans[1]

1 Consciousness, Cognition and Computation Group, Université Libre de Bruxelles, Brussels, Belgium, **2** Social Psychology Unit, Université Libre de Bruxelles, Brussels, Belgium, **3** Social and Developmental Psychology Department, University of Cambridge, Cambridge, United Kingdom

Abstract

The perspective that behavior is often driven by unconscious determinants has become widespread in social psychology. Bargh, Chen, and Burrows' (1996) famous study, in which participants unwittingly exposed to the stereotype of age walked slower when exiting the laboratory, was instrumental in defining this perspective. Here, we present two experiments aimed at replicating the original study. Despite the use of automated timing methods and a larger sample, our first experiment failed to show priming. Our second experiment was aimed at manipulating the beliefs of the experimenters: Half were led to think that participants would walk slower when primed congruently, and the other half was led to expect the opposite. Strikingly, we obtained a walking speed effect, but only when experimenters believed participants would indeed walk slower. This suggests that both priming and experimenters' expectations are instrumental in explaining the walking speed effect. Further, debriefing was suggestive of awareness of the primes. We conclude that unconscious behavioral priming is real, while real, involves mechanisms different from those typically assumed to cause the effect.

In this analysis, we used participants' walking speed as they entered the experiment room, (i.e., before priming) as a covariate. The results show no significant difference between the Prime (M = 6.270 SD = 2.15) and the No-Prime group (M=6.390 SD=1.11) in the time necessary to walk along the hallway after the priming manipulation (F (1, 119),1, g2 = .01).

Moving forward with projects

- Prototype + “Pilot A” + writeup
  - Run on 2-4 non-naïve subjects
  - Run analyses and publish replication report to Rpubs
  - Email report + link to staff for Pilot B approval
- “Pilot B”
  - Get class MTurk credentials
  - Run 2-4 naïve turkers (add comments box)
  - Email report to staff for final approval

Project power guidance

1. Choose key test and determine whether you can compute an ES
   - What test most directly captures the key interpretive claim of the study? (Often good to look at the abstract/intro)
2. If you can compute ES:
   - determine post-hoc power
   - Compute 80% power based on ES estimate from study
   - is this feasible in terms of price/time?
3. If you can’t compute ES
   - Talk to me/TAs
   - Consider if 2.5x is possible
   - Fall-back to original sample size

Preregistration

- Connect your github report to OSF
  - We will demo in class
- Get final signoff from me + TA
- Register!
- Collect data on agreed-upon sample
- Run scripts for planned analyses
- Then add exploratory analyses

Class MTurk guidelines

- Turkers are people! Treat them with respect
  - We aim for federal min wage – $7.25/hr
- Extensive piloting is intended to make experiments that are clear and do not fail
  - If there are failures, we pay for time spent, not work done
  - (Whenever possible, experiments should be designed to prevent fraud)
  - If you mess up, you will have to create a ”compensation HIT” to pay turkers for lost time

“catch trials”?

While watching the television, have you ever had a fatal heart attack?

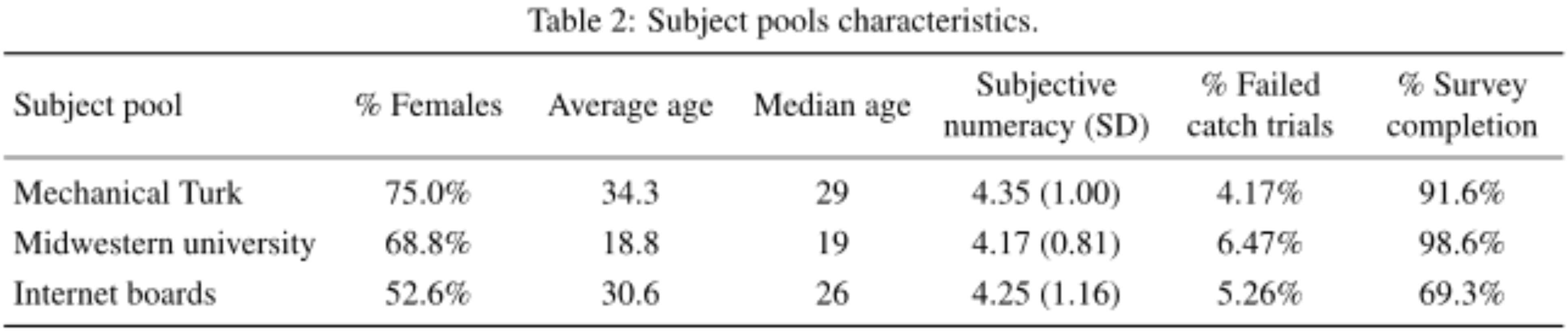


Table 2: Subject pools characteristics.

Subject pool	% Females	Average age	Median age	Subjective numeracy (SD)	% Failed catch trials	% Survey completion
Mechanical Turk	75.0%	34.3	29	4.35 (1.00)	4.17%	91.6%
Midwestern university	68.8%	18.8	19	4.17 (0.81)	6.47%	98.6%
Internet boards	52.6%	30.6	26	4.25 (1.16)	5.26%	69.3%

Paolacci et al. (2010)

Manipulation checks:

- Instructional manipulation checks
  - Tricky
  - Comprehension questions
- Experimental manipulation checks
  - Question about the stimulus (check viewing/retention of materials)
  - Question about the manipulation of interest
- Always critical to pre-set and pre-register your exclusion criteria
- If manipulation checks interact with experimental manipulation itself, can invalidate causal inference
  - E.g., excluding on “attention check” with a more attentionally demanding manipulation