Practical Best Practices in Psychological Science

SPSP Deep Dive Workshop

Alison Ledgerwood

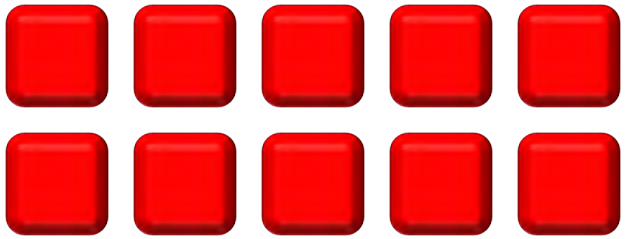
Feb 2020

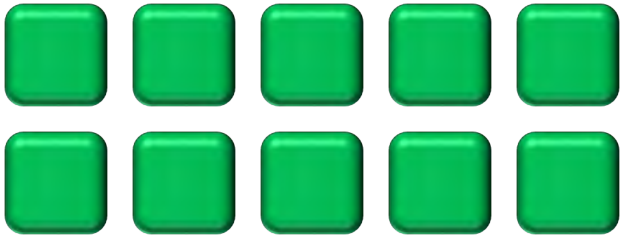
UCDAVIS
UNIVERSITY OF CALIFORNIA

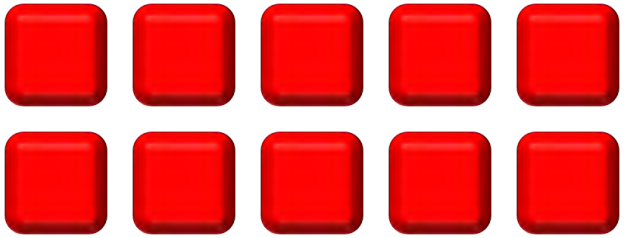
Calibrate your confidence

Part I: Understand Power 

Part II: Understand Type I error 

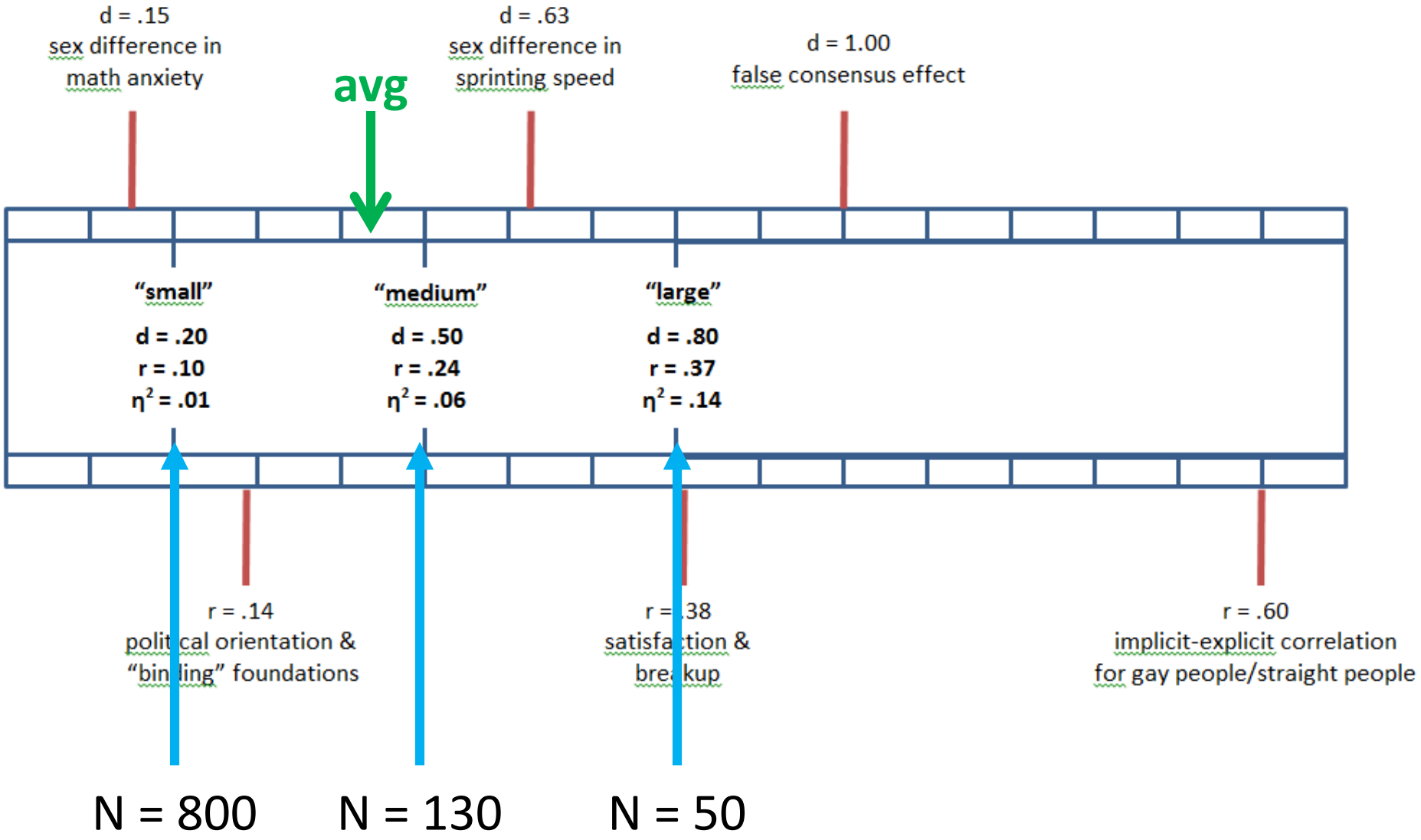






Part I: Understand Power

- Build your effect size & sample size intuitions
- Conduct power analyses
  - When planning a study
  - When assessing a study
- Understand how power works in interactions
- Understand other factors (beyond N) that influence power



*NOTE: This is for between-subject effects!

How to conduct a power analysis when planning a study

- SESOI: Smallest effect size of interest
  - BSSOW: Biggest sample size of worth-it-ness
    How much of my resource pie am I willing to spend to detect this effect?

How to conduct a power analysis when planning a study

- Formal *a priori* power analysis
  - With a GOOD effect size estimate (large sample, meta-analysis that properly accounts for pub bias)
    - G*Power, R, Westfall's PANGAEA for general ANOVA designs, Wang & Rhemtulla's pwrSEM for SEM
  - Or with a method that adjusts for publication bias and/or uncertainty
    - Perugini, Gallucci, & Costantini (2014) Safeguard power (uncertainty only)
    - McShane & Bockenholt's Power-calibrated effect size approach (uncertainty only)
    - Anderson, Kelley, & Maxwell (2017) BUCSS (both)

Exercise A: G*Power

- You're planning a study testing the correlation between prejudice (feeling thermometer) and cooperative behavior in a social dilemma (money returned in the Trust Game).
  - Effect size = ?
  - Meta-analysis or large studies with individual difference predictors of Trust Game behavior: r's ≈ .20
- Calculate sample size needed for 80% power to detect r = .20 in your study

Exercise B: G*Power

- You're planning a two-condition experiment and your best guess of the effect size is from a prior study's estimate for the difference between two conditions, $t(198) = 3.53$, $p < .001$, $d = .50$

- Use G*Power to calculate the total sample size (N) needed to have 80% power to detect this effect.

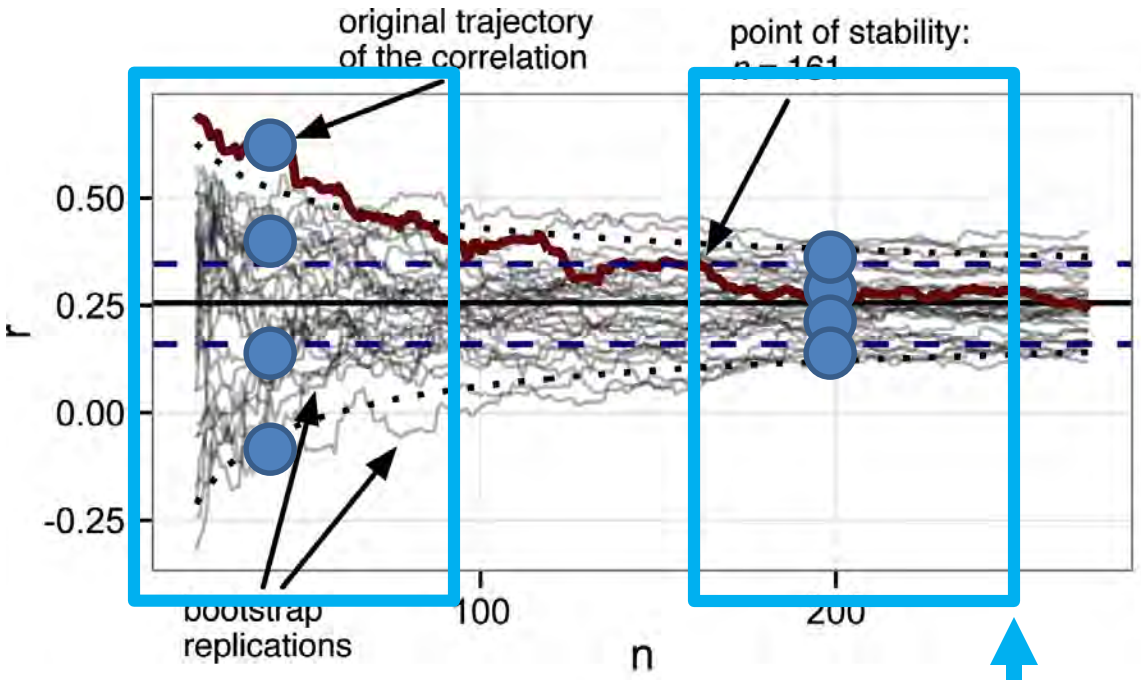
Exercise C: BUCSS

- You're planning a two-condition experiment and your best guess of the effect size is from a prior study's estimate for the difference between two conditions, $t(198) = 3.53$, $p < .001$, $d = .50$

- You DON'T think this study is big enough to precisely estimate the effect size and you DO think it may have been influenced by publication bias.

What's a "good" effect size estimate?

Low Uncertainty



At what sample size do correlations stabilize?

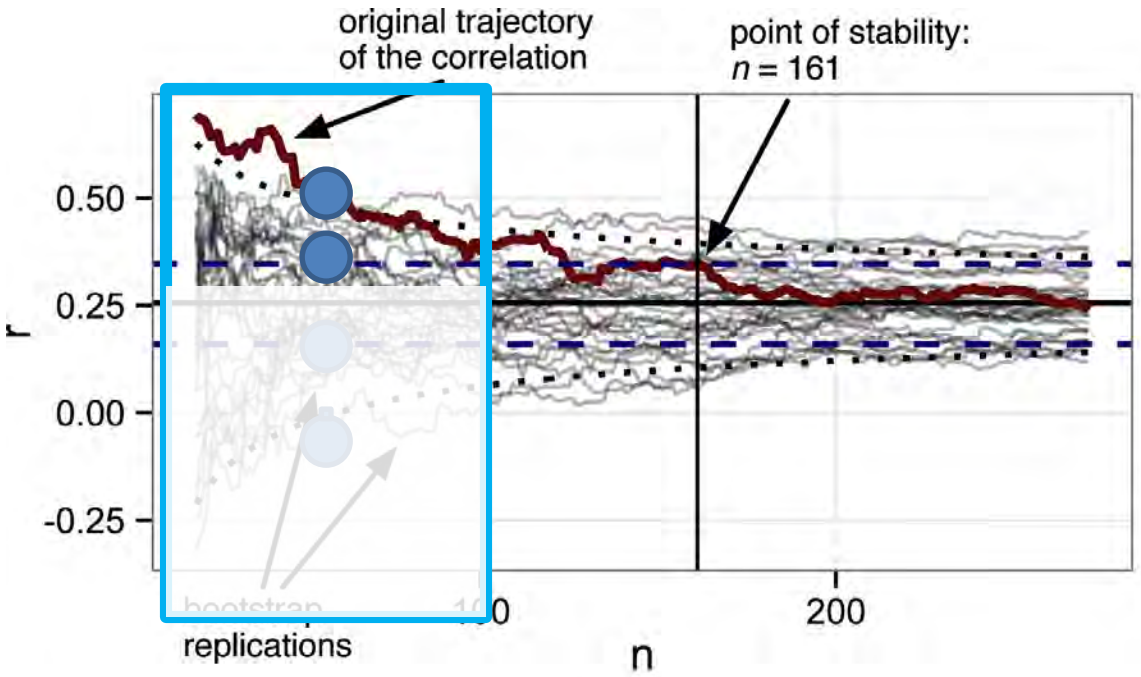
- Correlations stabilize with increasing *n*
- From which *n* on does a correlation stay within a *corridor of stability* (*COS*)?
- Depends on effect size, width of *COS*, and confidence level
- General suggestion: *n* should approach 250

Heuristic:
N = 250

Schonbrodt & Perugini, 2013

What's a "good" effect size estimate?

No Publication Bias



At what sample size do correlations stabilize?

- Correlations stabilize with increasing *n*
- From which *n* on does a correlation stay within a *corridor of stability* (*COS*)?
- Depends on effect size, width of *COS*, and confidence level
- General suggestion: *n* should approach 250

Schonbrodt & Perugini, 2013

Exercise C: BUCSS

- You're planning a two-condition experiment and your best guess of the effect size is from a prior study's estimate for the difference between two conditions, $t(198) = 3.53$, $p < .001$, $d = .50$
- You DON'T think this study is big enough to precisely estimate the effect size and you DO think it may have been influenced by publication bias.

Sample-Size Planning for More Accurate Statistical Power: A Method Adjusting Sample Effect Sizes for Publication Bias and Uncertainty

2017, Vol. 28(11) 1547–1562
© The Author(s) 2017
Reprints and permissions:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/0956797617723724
www.psychologicalscience.org/PS
SAGE

Samantha F. Anderson, Ken Kelley, and Scott E. Maxwell
University of Notre Dame

Exercise C: BUCSS

- You're planning a two-condition experiment and your best guess of the effect size is from a prior study's estimate for the difference between two conditions, $t(198) = 3.53$, $p < .001$, $d = .50$
- Use the Anderson et al. (2017) BUCSS Shiny Web App to calculate the **total sample size (N)** that you would need for your planned two-condition experiment if you wanted to account for publication bias (assume that studies are only published when $p < .05$) and if you want a 75% chance (assurance) that your own study will have 80% power.

designingexperiments.com → Web Apps → scroll down

Exercise C: BUCSS (cont'd)

- Now try to use the same Shiny App to calculate the N you would need if the prior study had the same effect size estimate but a smaller sample size—say, $t(98) = 2.48$, $p = .01$, $d = .50$. Note what happens. Play around with the assurance level.

Caveat

- These power analyses are appropriate IF the goal of your study is to determine the existence or direction of an effect
- Other goals:
  - Estimate the size of an effect (Schonbrodt & Perugini)
  - Equivalence testing (Lakens, 2017)

Anderson, Kelley, & Maxwell, (2017)

Sequential Analysis

- What if your range of plausible effect sizes yields a wide range of target Ns, or you want to conserve resources as much as possible?
- **Sequential analysis**
  - Lets you select *a priori* a total N (how large would you be willing to go if needed) and specific interim analysis points (where would you like to stop if you could), *without* inflating Type I error.
  - Calculate alpha cut-offs for each point.

Lakens (2014 EJSP)

Exercise D: Sequential analysis

- Using the total N you calculated in Exercise C, plan a sequential analysis that will allow you to peek at your data once halfway through data collection. What will your alpha cut-offs be for the interim analysis and final analysis?
  - Use this table (from Da Silva Frost & Ledgerwood, in press):

Divide your sample size into _ equal parts	Stop at		Alpha Threshold	Decision Guide	
	percent of total N	example (total N=600)			
2	50%	300	.025	p<.025?	if yes, significant. if no, continue collection
	100%	600	.034	p<.034?	if yes, significant if no, it is not significant
3	33%	200	.017	p<.017?	if yes, significant. if no, continue collection
	66%	400	.022	p<.022?	if yes, significant if no, continue collection
	100%	600	.028	p<.028?	if yes, significant if no, it is not significant

How to conduct a power analysis when assessing a study

- Sensitivity analysis
  - Goal: give people an intuitive sense of the power a study this size would have to detect effect size X.

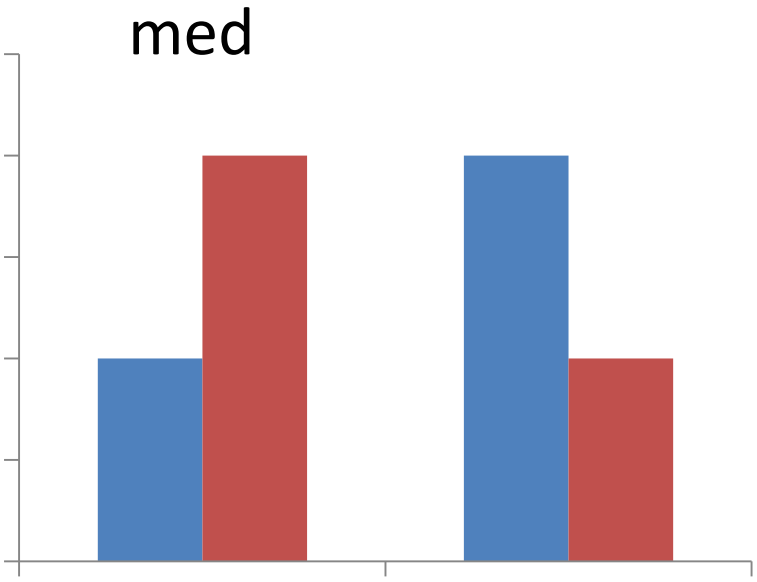
    Ex: In a two-group experiment, N = 140 would provide 80% power to detect an effect of Cohen's d = .48 and 60% power to detect an effect of Cohen's d = .38.
- ~~Post-hoc power analysis or "observed power"~~
  - Problem: You cannot use the effect size estimate from a study to compute that same study's power.
  - "It was sunny today, so chance of rain must have been 0%"

How to conduct a power analysis when assessing a study

- Sensitivity analysis
  - Goal: give people an intuitive sense of the power a study this size would have to detect effect size X.

    Ex: In a two-group experiment, N = 140 would provide 80% power to detect an effect of Cohen's d = .48 and 60% power to detect an effect of Cohen's d = .38.
- ~~Post-hoc power analysis or "observed power"~~
- Use your effect size intuitions

Crossover

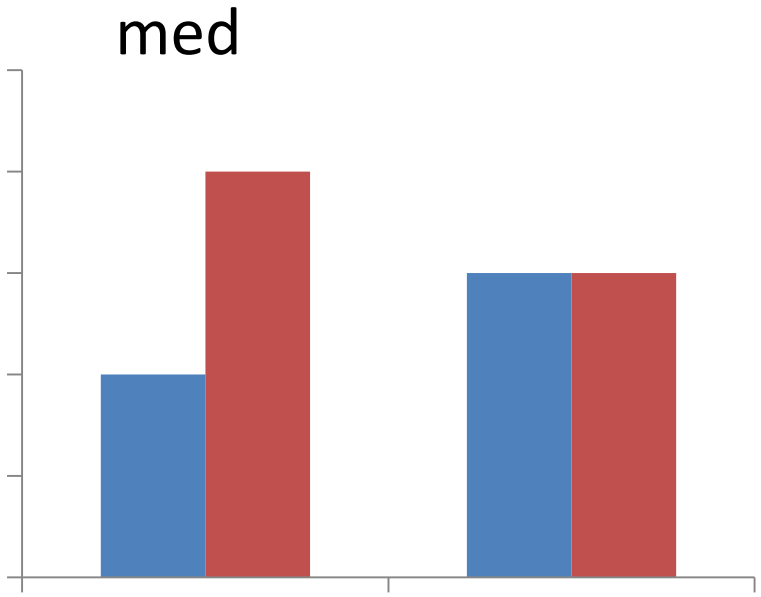


N = 130 +N = 130

N = 260

(to power follow-up comparisons)

Knockout



N = 130

N = 520

50% Attenuation: $N_2 = 14*N_1 = 1820$

Giner-Sorolla, 2018; Ledgerwood, 2019

Beyond N: Other ways to boost power

- Within-subjects designs
- Increase the reliability of your measures
- Increase the strength of your manipulation
- For experiments: Select *a priori* a covariate that should correlate strongly with your DV
- Collaborate across labs and aggregate the results (StudySwap, Psych Science Accelerator)

Calibrate your confidence

Part I: Understand Power 

Part II: Understand Type I error 

Part II: Understand Type I error rate

- How do various research practices influence your Type I error rate?
- Distinguish data-independent from data-dependent analyses
  - Pre-analysis plans
- Think critically about the broader concept of preregistration
  - Different goals of preregistration and how to achieve them

Practices that inflate Type I error

- Arrange these practices in order of increasing Type I error rate:

~7% (1) Running an analysis with and without a covariate

~6% (2) Collecting N = 100, getting p = .09, and deciding to continue data collection to N = 200

14% (3) Running a preregistered 2x2 ANOVA & interpreting the main effects and interaction

can approach 100% (4) Testing for incremental validity [X predicts Y over and above C] using multiple regression in a large sample.

Covariates: See Wang, Sparks, Gonzales, Hess, & Ledgerwood, 2017

Optional stopping: See Sagarin, Ambler, & Lee, 2014

Type I error rates in factorial ANOVAs: See Cramer et al. 2016

How to test for incremental validity the right way: See Wang & Eastwick, in press

Pre-analysis Plans

- Exploring your data is very important, but in order to calibrate your confidence, you need to know when your Type I error rate is inflating
- Pre-analysis plans can help distinguish data-independent from data-dependent analyses
  - Data-independent analyses: Interpret *p*-values as diagnostic of likelihood of result
  - Data-dependent analyses: Ignore *p*-values or interpret more tentatively

Pre-analysis Plans

- Identify planned analyses
- Constrain foreseeable researcher decisions for those analyses (sometimes you can't; e.g., surprise skew)
- Be clear and precise so that you can tell when you're actually doing something data-dependent! (e.g., vague vs. specific exclusion criteria)
- Options (each has pros and cons)
  - Internal: For yourself or for your lab
  - AsPredicted.org: Useful template for experiments
  - OSF: Unstructured, upload anything

Pre-analysis Plans: AsPredicted

AsPredicted Questions

(version 2.00)

3) Dependent variable. Describe the key dependent variable(s) specifying how they will be measured.

Example: Simple average GPA across all courses during the first semester after the intervention.

4) Conditions. How many and which conditions will participants be assigned to?

Example 1: Two conditions: Offering summer program: yes vs no.
Example 2: 12 conditions in a mixed design lab study. Participants will be assigned to one of four conditions: math training, verbal training, memory task, or control (4 between-subject conditions). Each participant will complete a math test, a verbal test, and a memory test (3 within-subject conditions).

5) Analyses. Specify exactly which analyses you will conduct to examine the main question/hypothesis.

Example. Linear regression predicting the simple average GPA in the semester after the intervention with a dummy variable indicating whether the participant was offered the summer program or not (intention-to-treat analysis). We will also conduct the same regression controlling for simple average GPA during the semester before the intervention, gender, & household income (an 8-point scale ranging from 1 = below $20,000 and 8 = above $150,000).

6) Outliers and Exclusions. Describe exactly how outliers will be defined and handled, and your precise rule(s) for excluding observations.

Example 1: We will compute the overall mean and standard deviation across all conditions, and winsorize at 2.5 SD above/below the mean.
Example 2: We will exclude participants who incorrectly answer at least 2 of our 3 attention check questions.
Example 3. We will exclude any participants who complete the survey in less than 30 seconds.

7) Sample Size. How many observations will be collected or what will determine sample size?
No need to justify decision, but be precise about exactly how the number will be determined.

Preregistration Flavors

Definition	Goal
Pre-analysis plan	Distinguish data-independent vs. data-dependent analyses
Write down as much information as you can about your study	Transparency: Someone else can check what you said you planned ahead of time against what you actually wrote down
Record your theoretical predictions	Theory falsification
Record your intuitive predictions	Figure out how good you are at guessing?
Record the existence of your study	Combat publication bias
Registered report	All of the above plus reviewer objectivity

See Ledgerwood (2018 PNAS); Ledgerwood & Sakaluk SIPS presentation

Exercise E: Pre-analysis plan

- Take 10 minutes to create a pre-analysis plan for a recent or planned study.
- Then, swap with your neighbor
- Neighbor: Try to poke holes in it. Any researcher decisions that haven't been anticipated or fully constrained?

Exercise F: Preregistered prediction

- Take 5 minutes to write down the theoretical predictions for a recent or planned study.
- Then, swap with your neighbor
- Neighbor: Try to poke holes in it. Is a pattern of results specified that would *reduce* confidence in the theoretical prediction?

Questions/Discussion