

# Running Experiments

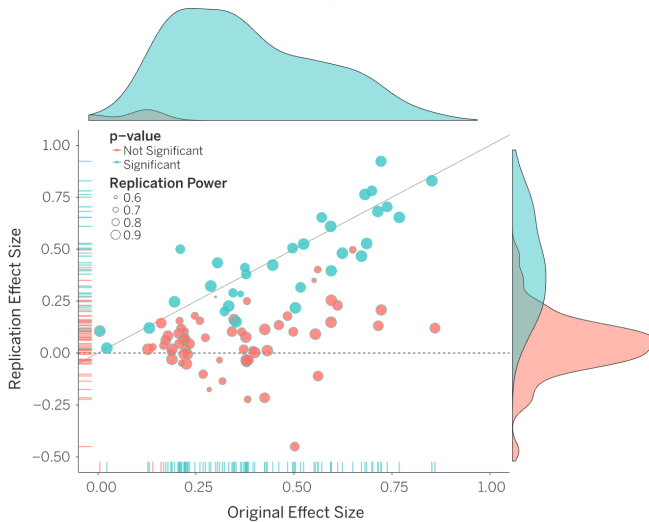
Kevin Dorst (kmdorst@mit.edu)

24.805, Fall 2025

## I. Why all the kerfuffle?

Replication crisis (Open Science Collaboration 2015):

This is why we've moved toward  
 (1) statistical rigor,  
 (2) pre-registration, and  
 (3) data transparency



**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.

## II. Steps to running an experiment

- o) Learning basics of statistics (conceptually).
  - 1) IRB (institutional review board) approval.
  - 2) Getting funding.
  - 3) Hypothesis, generated via theory (and perhaps simulations).
  - 4) Making your survey: [Qualtrics](#).
  - 5) Recruiting subjects: [Prolific](#).
  - 6) Piloting and troubleshooting.
  - 7) Analyzing pilot results: [R and RStudio](#) + ChatGPT
  - 8) Pre-registration: [AsPredicted.com](#)
  - 9) Run and analyze study; create visualizations.
  - 10) Post (anonymized) data and R scripts on [ResearchBox](#)
  - 11) Repeat. Get better. Do science.

MIT IRB = COUHES: Committee on the Use of Humans as Experimental Subjects.  
 ≈ \$500-1000 per exp (start to finish)

Syncs with AsPredicted

And find a lab meeting to attend!  
 The best way to learn is with people.

### o) Learning Statistics

No calculus (and few Psets) needed; goal is to understand *conceptually* so that you can wrangle ChatGPT to write code/models for you.

Options:

Statistics help and office hours:

- 1) [IQSS at Harvard](#)
- 2) [Statistical software help at MIT](#)
- 3) ? Maybe [Data@MIT](#)?

- Bayesian, no calculus: [McElreath 2018, Statistical Rethinking](#).
- Bayesian, some calculus: [MIT online course](#)
- Frequentist (significance testing, confidence intervals, regressions), no calculus: classes 17–26 of [MIT course](#)

In all cases, try to learn it *with an LLM*. They are very good tutors for this sort of thing.

### 1) IRB approval

Training. At MIT you need a faculty sponsor (probably me or Justin, or a linguist); but you can be the PI.

Most things philosophers want to do have ‘exempt’ status at MIT.

McElreath [Youtube lecture series](#)

Often easier to have a particular goal (‘I want to do a significance test for *this* hypothesis’) to get yourself to learn what you need to learn.

[Show my application]

### 2) Getting funding

Annoying, as a grad student. Talk to me/Kieran/Jen.

MIT should have some internal funds for this.

I’m happy to provide some (supply-constrained!) startup funds.

### 3) Generating a hypothesis

Start with theory. Use to make general hypothesis, and extract from that testable predictions.

Testable predictions should involve manipulating at least one *dependent variable* (DV), and then measuring its effect on at least one *independent variable* (IV).

**Running example:** Does clarity eliminate (b/t-subject) hindsight bias?

*Hypothesis:* ambiguity is necessary for hindsight bias.

*Testable prediction:* measure hindsight bias using between-subjects test. In ambiguous scenarios, informed subjects should have higher estimates than uninformed subjects; in clear scenarios, no difference.

→ Often hard to test ‘no difference’ claims. So extract qualitative prediction: clear scenarios should lead to *less* HB than ambiguous ones.

Simulations are helpful here, especially if your theory is Bayesian.

Often interested in interaction effects between multiple DVs and IVs.

Informed vs. uninformed subjects; ‘How likely would you have thought *q* was?’

There should be an *interaction effect* between one DV (being informed vs not) and another (having ambiguity vs. clarity)

### 4) Making Survey

Use Qualtrics. Most tedious part. Use ChatGPT to help write materials.

Consent; Prolific code; demographics/attention-check; instructions and comprehension checks; questions (question names!!); survey flow.

Take the test yourself! Revise; retake.

Get account through MIT.

## 5) Recruiting subjects

Prolific best for subjects who pay attention / read instructions. (But it's expensive).

Alternatives: MTurk, Bovitz, Lucid.

Write blurb; update links (Qualtrics→Prolific; Prolific → Qualtrics); set screeners; exclude prior studies; estimate time / payment.

Best way to screen out subjects (I now know): automatically ask them to 'return' study if they fail a comprehension check.

Pay whoever doesn't return it (not worth the hassle.)

Make sure it's right at the beginning, and they have two chances to answer it right. [Prolific policies](#).

## 6) Piloting

Important!! Even once you're good at writing materials, almost never will you write things correctly the first time. Must test it.

It's sensible to pilot before pre-registering. The core role of pre-registering is for the main study, for internal validity/replicability. Expect, at the start or with difficult surveys, to do lots of piloting.

*But* be careful: if you run a million pilots and only find it working under very specific conditions [eg adjusting your stimuli just so], that's a reason to worry about robustness

Include open-ended feedback form, 'Was there anything about this survey that you found especially confusing?'

Omit in final study (for time/\$\$).

Make sure to get at least 50–100 subjects (depending on test), so that you're not chasing noise.

Maybe new alternative: use LLMs??  
<https://www.expectedparrot.com/>.

Study (/pilot) launch checklist:

- [If full study:] Have you pre-registered it?
- Have you re-named your questions?
- Have you updated the Prolific link to Qualtrics?
- Have you updated the Qualtrics link back to Prolific?
- Have you previewed the survey, and reset counters after preview?
- Have you deleted the data from your previews?
- Have you published the final version of Qualtrics?

## 7) Analyzing pilot results

RStudio, with help of ChatGPT.

Export raw data from Qualtrics. Import to RStudio.

Data wrangling to 'long format' (ChatGPT!), grabbing DVs and IVs.

Exclude those who fail; track comprehension/attention pass rates.

Descriptive stats; simple visualizations.

Simple statistics as sanity checks (t-tests; bootstrapped CIs).

Usually there will be things you don't expect. Look at feedback, data, descriptive stats and visualizations, to get a sense for what might've happened.

Formulate and test the official stats/model you're going to pre-register.  
→ ChatGPT can help you brainstorm what regressions/statistical tests to run (It's pretty good at that sort of thing). Try to keep models simple.

If you want to be careful (eg if expensive), simulate data that you expect, and check that your model recovers the true parameters.

Make sure to test your proposed statistical model on your pilot data!

- If it doesn't converge, maybe your model is too complicated.
- Can give you a sense of how much bigger sample size you'll need for confident estimates.

## 8) Pre-registration

I use [AsPredicted.com](https://aspredicted.com) (it's simpler).

This study: <https://aspredicted.org/3s3g-tqrf.pdf>

(Get some examples; have ChatGPT help you write it.)

## 9) Run and analyze it!

This should be quick (but nerve-wracking!!)—you should already have your R script completely written, so just import it and go!

First look at simple stats and visualizations. Then run models.

## 10) Post data and R scripts online

On [ResearchBox](https://www.researchbox.org/), probably

## 11) Keep at it!

Experiments will fail. Even if your hypothesis is true, pilots and even full runs won't work the way you expect. That's okay! Keep at it.

There are a lot of skills to master. But once you get them down, you'll have a superpower compared to other philosophers.

Interpretability usually beats statistical propriety.

Often full experiment needs to be 4–6 times as large as pilot for sufficient statistical power. Model confidence often scales roughly by  $\sqrt{n}$ . So if go from 100 to 400 subjects, credible intervals probably shrink by factor of 2.

You can also use OSF

Save their outputs so you don't need to re-run big statistical models each time

It gets better with practice.