

Experiments

Manoel Horta Ribeiro
manoel@cs.princeton.edu

Recap: Experiments Rock

- We show we can estimate the ATE :

$$\hat{\tau} = \frac{1}{|\mathcal{D}_1|} \sum_{(i,r): T_{ir}=1} Y_{ir} - \frac{1}{|\mathcal{D}_0|} \sum_{(i,r): T_{ir}=0} Y_{ir}$$

- Exchangeability:

$$(Y_{ir}^0, Y_{ir}^1) \perp T_{ir}.$$

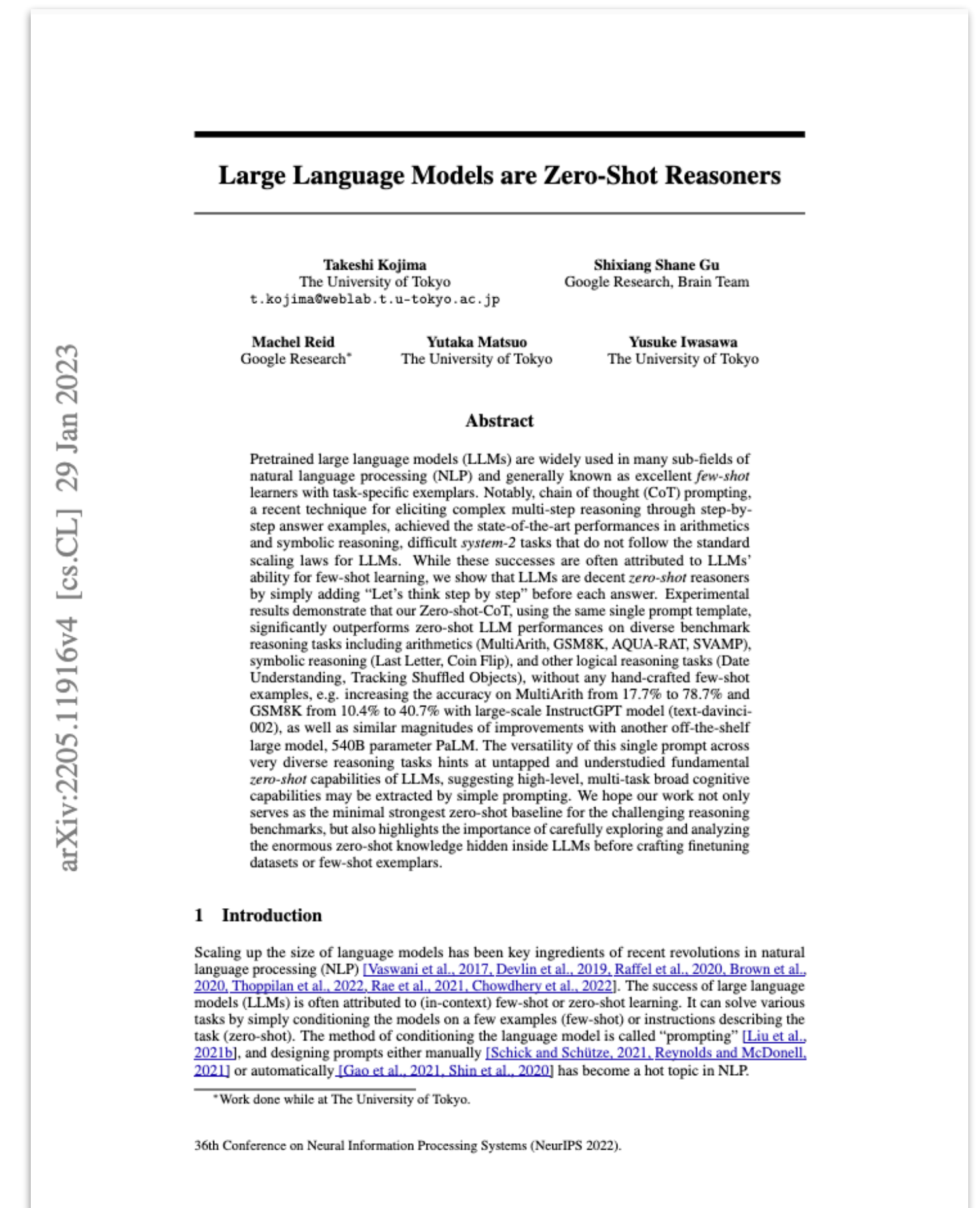
**Today's lecture is about “what
can go wrong” in an experiment!**

Chain of Thought (CoT)

- In an influential paper, Kojima et al. (2022) proposed changing the way we prompt models.
- **Hypothesis:** Asking the model to reason improves performance.

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there? A: The answer (arabic numerals) is

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there? **A: Let's think step by step.**



Potential Problem #1:
You measure the wrong thing

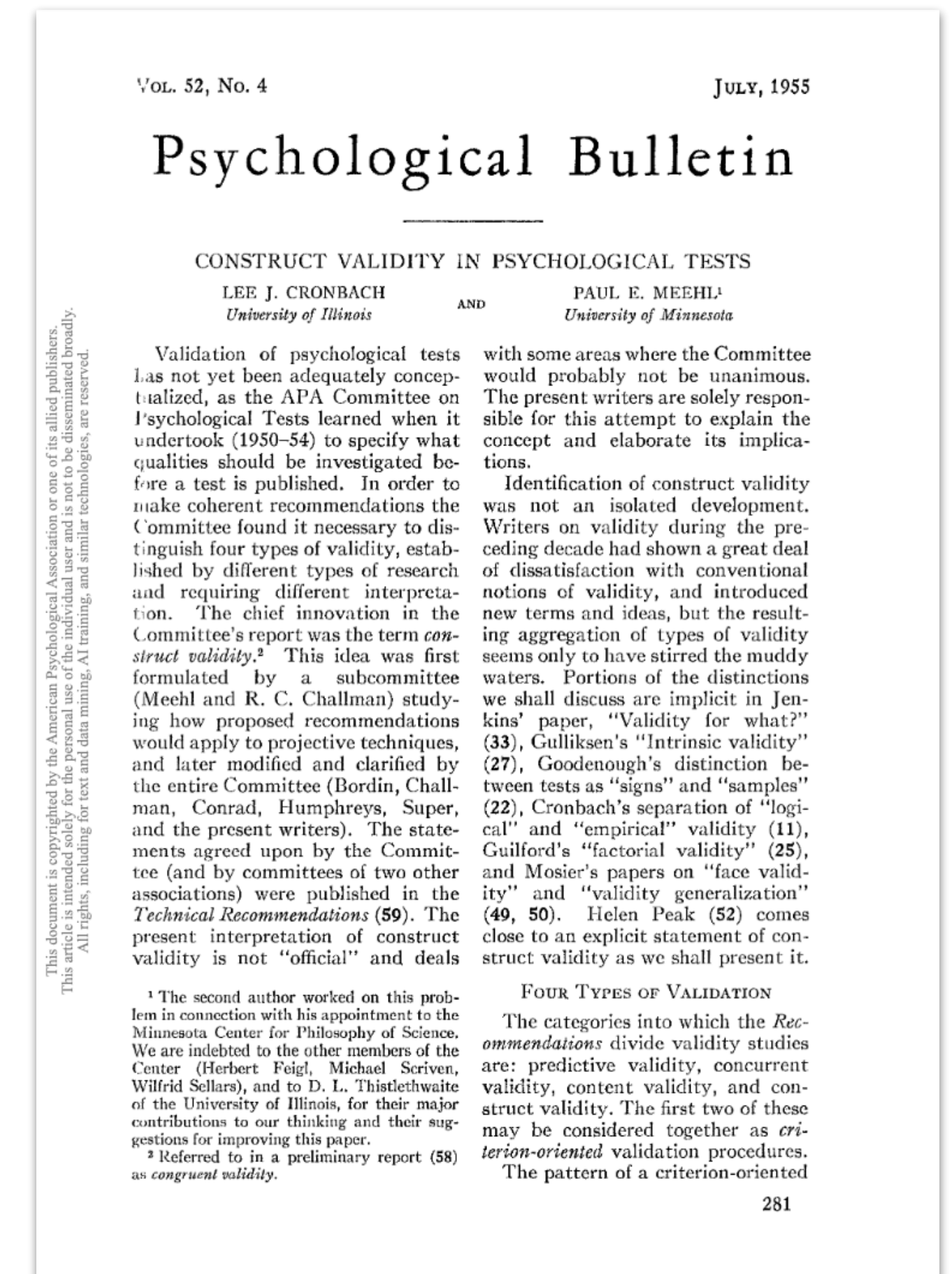
Construct vs. Measurement

- In many cases, we care about an abstract *construct*.
 - E.g., we care about LLMs mathematical reasoning capabilities.
 - This cannot be “operationally defined,” like height or temperature.
- The process of turning this construct into a concrete measurement is called *operationalization*.
- When we compare LLMs’ capacity to solve math problems in specific benchmarks, we get a concrete number.

Construct Validity

The extent to which a test or measurement accurately represents the concept or construct it intends to measure.

- How can we assert construct validity?
 - There's no statistic for construct validity!
 - We need to build an argument using multiple sources of evidence.



Construct Validity (cont'd)

- Does the measurement *span the construct's* domain?
 - If we only evaluate on algebra word problems, we are not measuring math reasoning broadly.
 - If CoT helps algebra but hurts geometry, can we say that CoT improves math reasoning?
- Does the measurement predict external criteria of interest?
 - If we give two models (CoT vs. Regular) to people doing mathematics research, do they prefer outputs from the CoT version?

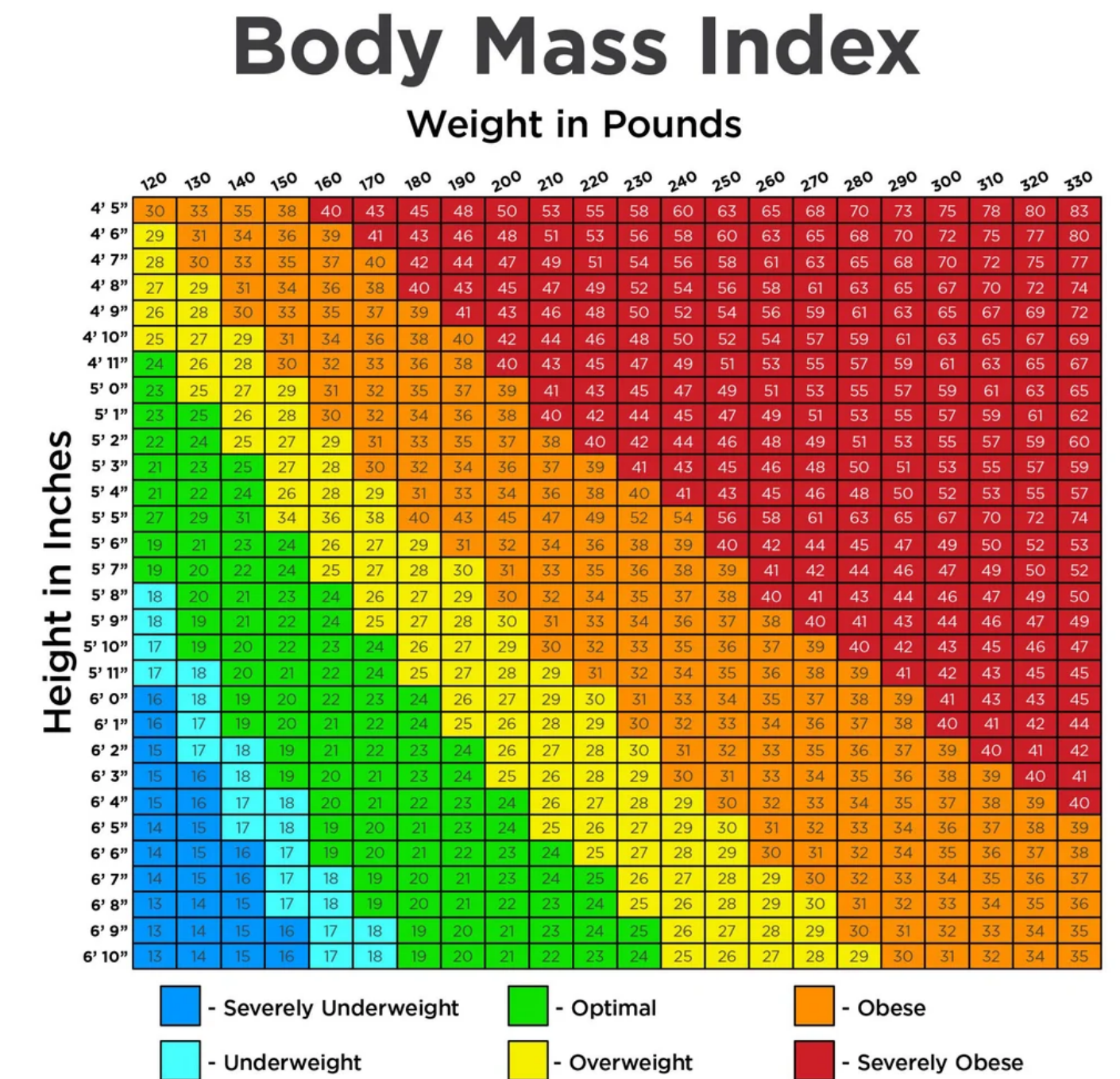
Construct Validity (cont'd)

- Does the measurement *align with other measures of the same construct*?
 - Do the CoT findings generalize across very different datasets related to math problems?
- Is the measurement *driven by a different construct*?
 - Maybe the model isn't reasoning at all, it is just that when there are more tokens in the context, the model is more likely to give a correct answer?

Reliability

The extent to which a test or measurement produces consistent results when the underlying object being measured is the same.

- BMI is a valid measure at the population level.
- However, it is unreliable at the individual level.
- Again, there's no single statistic to ensure reliability!



Reliability (cont'd)

- Do we get the same result if we repeat the measurement?
 - Run the model multiple times with different random seeds.
- Do equivalent measurements give similar results?
 - Create two matched math test sets, measure CoT effect in both.
- Do items intended to measure the same construct agree?
 - Break the dataset into subskills. Measure CoT gains across them.

Potential Problem #2:

You don't detect an effect when
the effect does exist

Thought Experiment



- Suppose the coin is unfair.
- How often will a statistical test tell you it is unfair?
- With low n , probably very rarely.

Hypothesis Testing (Recap)

- Assume the null hypothesis H_0 is correct.
- Measure p , the probability of observing a test statistic as extreme or more extreme than observed, given H_0 .

$$p = P(|Z| \geq |z_{obs}| \mid H_0)$$

- If $p \leq \alpha$ we reject the null hypothesis and assume that there is an effect. Typically $\alpha = 0.05$.

Power

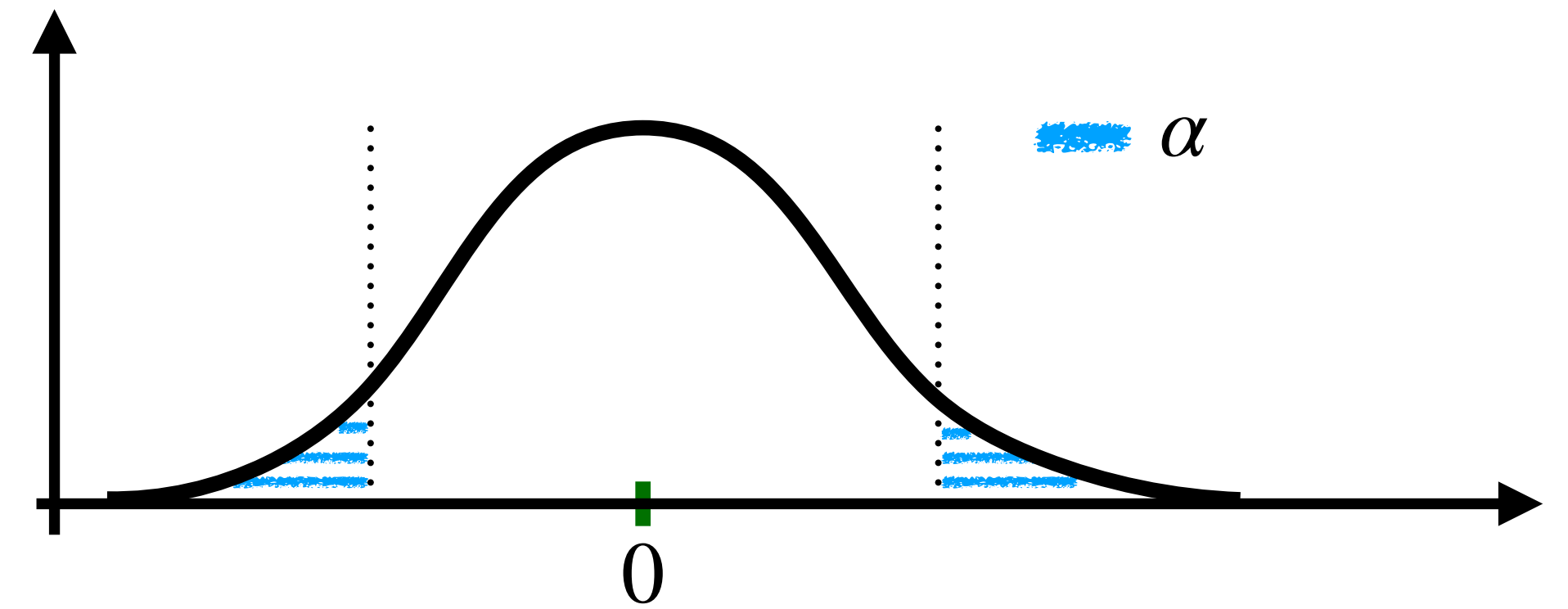
- But what if H_0 it is false and H_1 is true?
- Not all experiments can reject the null very often!
- Let β be the probability of “missing out” on an effect.
 - $1 - \beta$ = the probability we reject the null when it is false.
 - We call $1 - \beta$ the statistical power of an experiment.

Power

- Assuming the null, we have a distribution of Z :

$$Z = \frac{\hat{\tau} - \tau}{SE(\hat{\tau})} = \frac{\hat{\tau} - 0}{SE(\hat{\tau})} \sim \mathcal{N}(0,1)$$

- A test rejects the null if the probability of this happening under the null is unlikely.

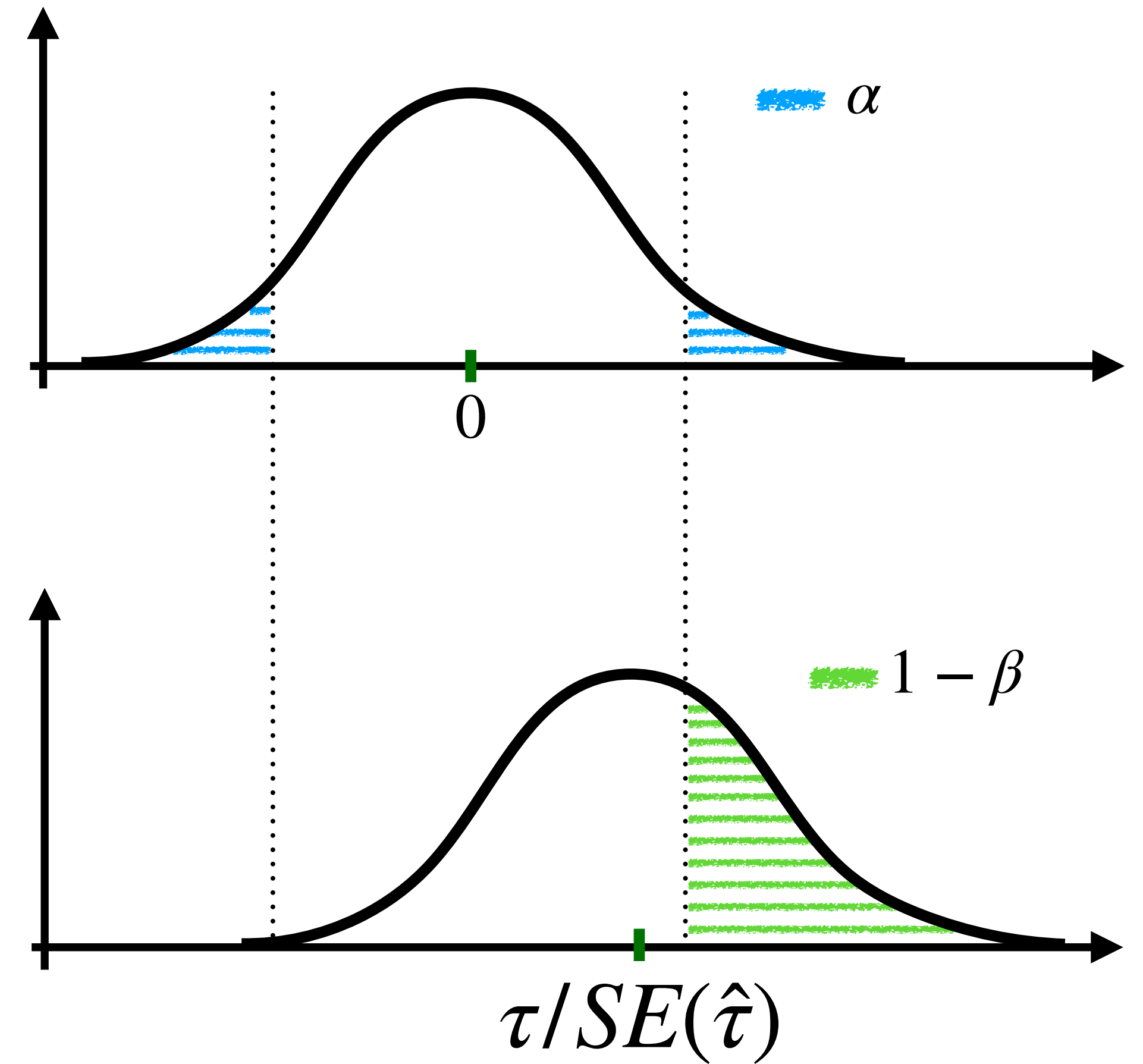


Power

- Now, assume that there is an effect! Then the distribution of Z under the null hypothesis is:

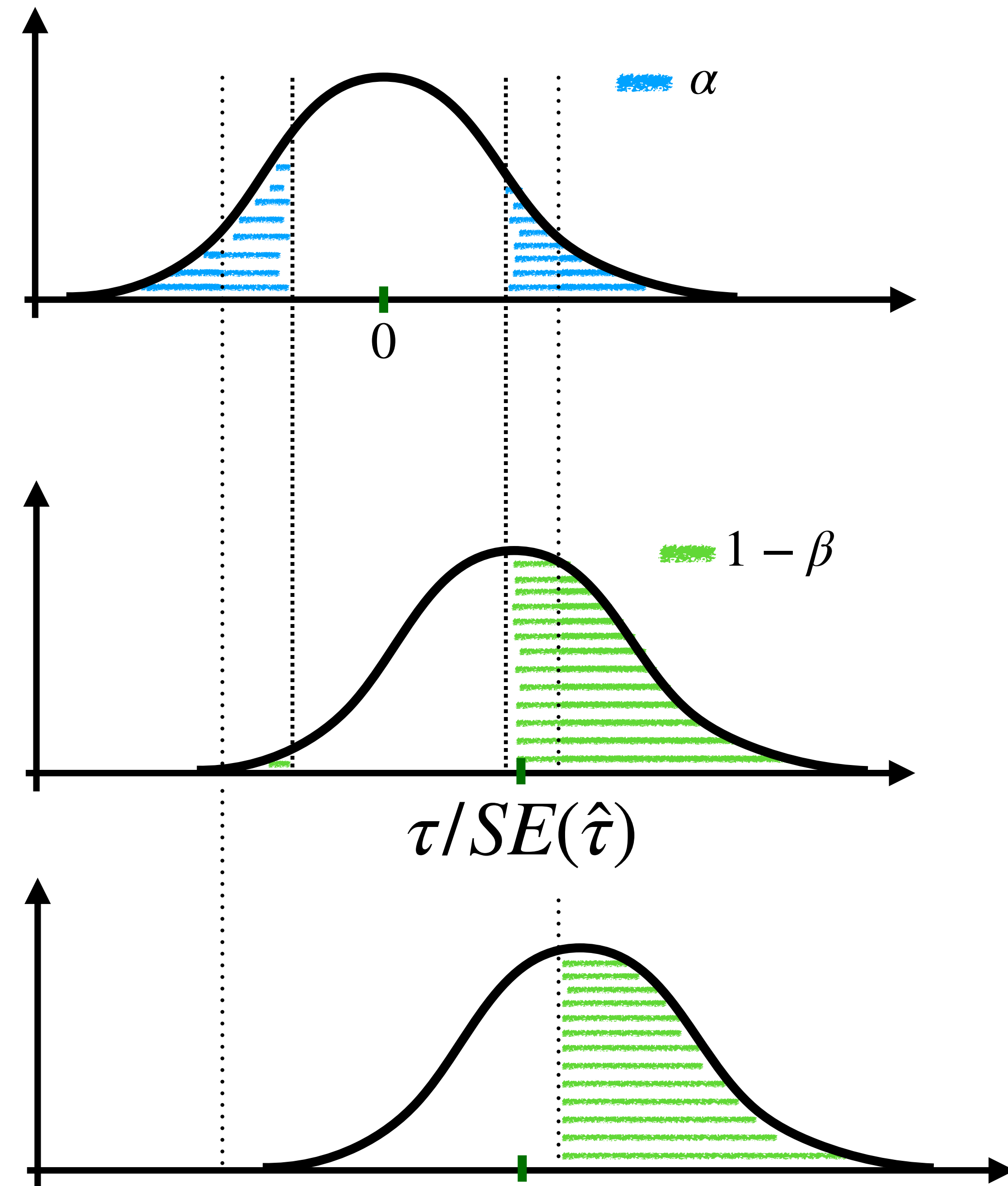
$$Z = \frac{\hat{\tau} - \tau}{SE(\hat{\tau})} = \frac{\hat{\tau} - 0}{SE(\hat{\tau})} \sim \mathcal{N}\left(\frac{\tau}{SE(\hat{\tau})}, 1\right)$$

- Power:** green area / total area



Increasing power

- Power depends on:
 - $Var(\hat{\tau})$, because $SE(\hat{\tau}) \sim \sqrt{Var(\hat{\tau})}$
 - n , because $SE(\hat{\tau}) \sim \sqrt{1/n}$
 - The significance threshold (α)
- To increase power, we can increase n or decrease the variance!



Power for the Z-test

Under the alternative hypothesis,

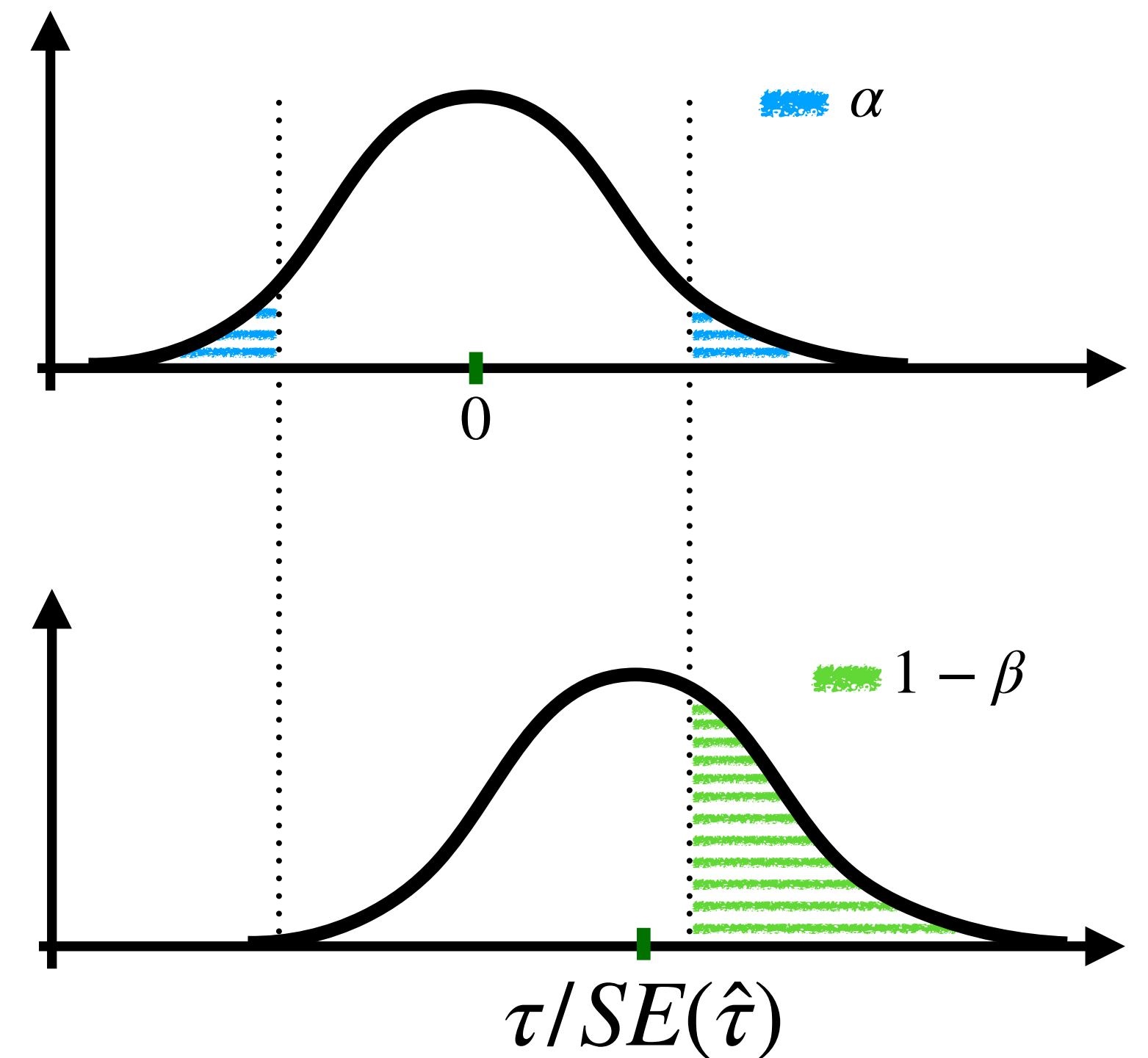
$$Z \sim \mathcal{N}(\tau/SE(\hat{\tau}), 1)$$

Then, we have that power equals to:

$$1 - \beta =$$

$$\Pr(|Z| > z_{1-\alpha/2} \mid H_1) =$$

$$1 - \Phi(z_{1-\alpha/2} - \tau/SE(\hat{\tau})) + \Phi(-z_{1-\alpha/2} - \tau/SE(\hat{\tau}))$$



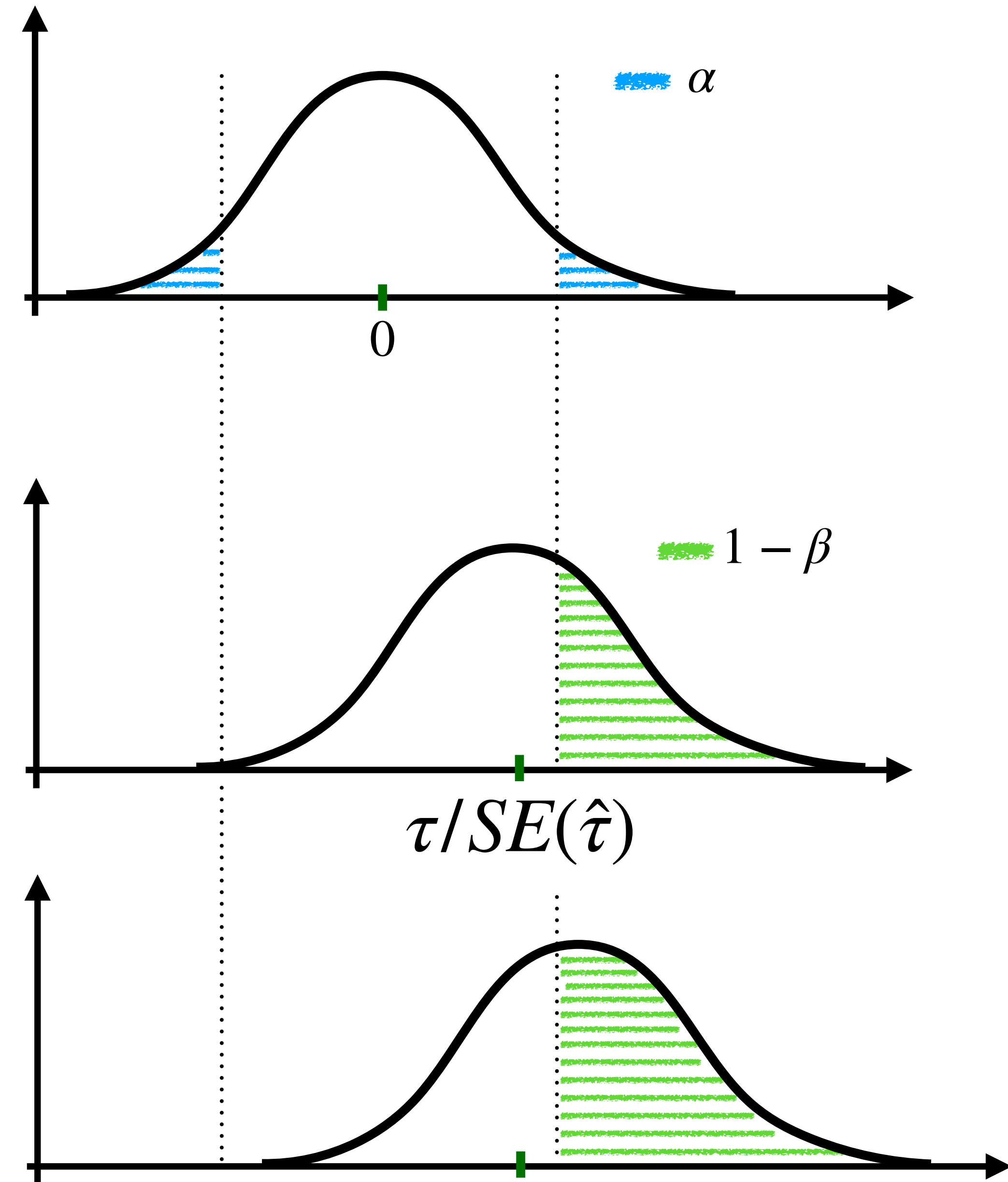
Power simulations

- In many cases, the estimator has no closed-form variance.
- The design of the experiment is complex (has clustering, blocking, and noncompliance)
- In this case, it is common to simulate power.
 - Specify a data-generating process, an effect size, a sample size, and noise.
 - Simulate data from this simulated experiment, and estimate the power empirically.

Minimum Detectable Effect

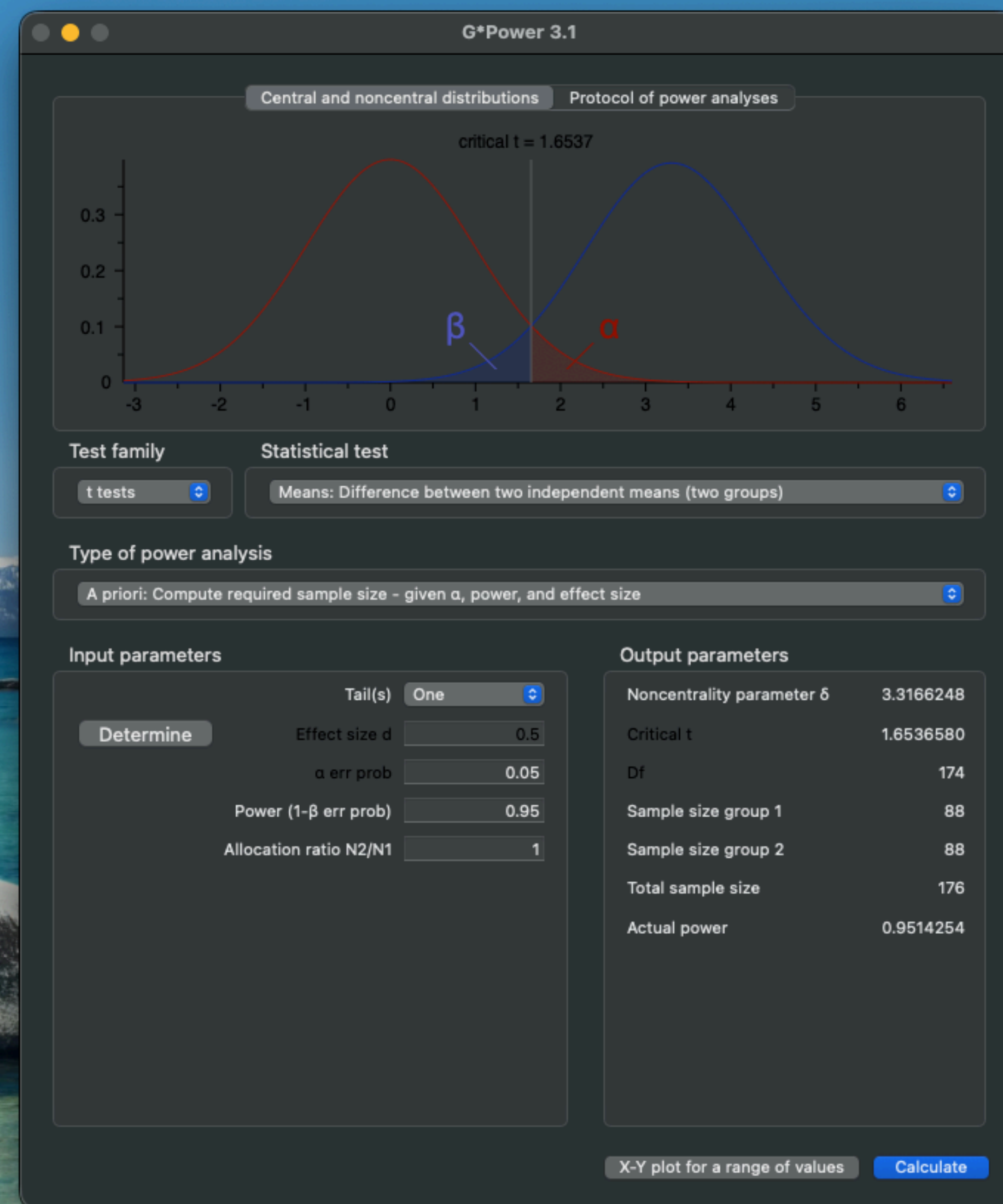
“What is the minimum effect size my experiment is high powered enough to reliably detect”

$$MDE = \left(z_{1-\alpha/2} + z_{1-\beta} \right) \frac{\sigma}{\sqrt{n}}$$



Experimental Design to Increase Power

- “Blocked randomization”
 - **Headache treatment:** Same demographics in both treatment and control groups.
 - **CoT example:** matched questions in treatment and control group (algebra, geometry, etc).
- “Within subjects”
 - **Headache treatment:** Each participant takes a placebo once and the actual medicine once.
 - **CoT example:** having multiple runs for the same question.



Potential Problem #3:

You detect an effect where the
effect does not exist

Thought Experiment

- Suppose CoT prompting does not work.
- What if we tested 20 different prompts?
 - Let's think step by step.
 - Describe your thinking step by step.
 - Write your thought process out loud as you solve the problem.
 - ...
- For each prompt, there's 5% of chance we reject the null
 - This is completely orthogonal to the sample size!

Fooling *Others* with Statistics

- If you keep testing hypotheses until something is statistically significant, you will likely get a false positive!
- Huge problem for science! If we only report statistically significant results, many findings will be false positives.



Correcting for multiple comparisons

- Let $H_1, \dots, H_m$ be a family of null hypotheses
- Let $\mathcal{J}_0$ be the set of indices where the null is actually true.
- Let $p_1, \dots, p_m$ be their corresponding p -values
- Family-wise error rate (FWER) = probability of rejecting at least one true null hypothesis (i.e., getting a false positive).

Bonferroni Correction

- **Idea:** reject H_i if $p_i \leq \alpha/m$

$$\text{FWER} = \Pr \left(\bigcup_{i \in \mathcal{J}_0} p_i \leq \frac{\alpha}{m} \right) \leq \sum_{i \in \mathcal{J}_0} \Pr(p_i \leq \frac{\alpha}{m}) \leq |\mathcal{J}_0| \frac{\alpha}{m} \leq \alpha$$

- If we do that, the FWER is α !
- If we test 20 prompts, we will obtain a FWER = 0.05, $\alpha = 0.0025$!

Other correction schemes

- Bonferroni correction is fairly conservative!
- One less conservative approach is Holm-Bonferroni
 - Order your p -values from bigger to smaller;
 - Test the i th smallest p -value with $\alpha/(m - i + 1)$.
 - “Spend” your error budget on the most convincing result.”
- Hierarchical testing:
 - test one primary endpoint first;
 - If significant, “unlock” secondary endpoints in a pre-set order.

When should you correct?

- **When to correct?** Correct when you will (or could reasonably) claim discovery from *any* of several tests.
- One way to avoid losing power: label some hypotheses as confirmatory, others as exploratory (pattern finding).
- The exact rituals around correcting are somewhat disputed.

Problem solved?

- Researchers can “ p -hack,” consciously or not in many ways.
- Changing data processing steps.
- Changing model specifications.

“A researcher when faced with multiple reasonable measures can reason (perhaps correctly) that the one that produces a significant result is more likely to be the least noisy measure, but then decide (incorrectly) to draw inferences based on that one only.”

Fishing, Commitment, and Communication: A Proposal for Comprehensive Nonbinding Research Registration

Macartan Humphreys

*Columbia University, Department of Political Science, 7th floor, IAB Building,
420 West 118th St., New York, NY 10027
e-mail: mh2245@columbia.edu (corresponding author)*

Raul Sanchez de la Sierra

*Columbia University, Department of Economics, 1022 IAB Building,
420 West 118th St., New York, NY 10027
e-mail: rs2861@columbia.edu*

Peter van der Windt

*Columbia University, Department of Political Science, 7th floor, IAB Building,
420 West 118th St., New York, NY 10027
e-mail: pv2160@columbia.edu*

Edited by R. Michael Alvarez

Social scientists generally enjoy substantial latitude in selecting measures and models for hypothesis testing. Coupled with publication and related biases, this latitude raises the concern that researchers may intentionally or unintentionally select models that yield positive findings, leading to an unreliable body of published research. To combat this “fishing” problem in medical studies, leading journals now require pre-registration of designs that emphasize the prior identification of dependent and independent variables. However, we demonstrate here that even with this level of advanced specification, the scope for fishing is considerable when there is latitude over selection of covariates, subgroups, and other elements of an analysis plan. These concerns could be addressed through the use of a form of comprehensive registration. We experiment with such an approach in the context of an ongoing field experiment for which we drafted a complete “mock report” of findings using fake data on treatment assignment. We describe the advantages and disadvantages of this form of registration and propose that a *comprehensive but nonbinding* approach be adopted as a first step to combat fishing by social scientists. Likely effects of comprehensive but nonbinding registration are discussed, the principal advantage being communication rather than commitment, in particular that it generates a clear distinction between exploratory analyses and genuine tests.

1 Introduction

There is a growing concern regarding reporting and publication bias in experimental and observational work in social science arising from the intentional or unintentional practice of data fishing. The adoption of registries provides one possible response to the problem, but while the potential benefits of preregistration of research designs are easily grasped there has been essentially no adoption of the practice by political scientists. Moreover, even with agreement on registration

Authors' note: Our thanks to Ali Cirone, Andy Gelman, Grant Gordon, Alan Jacobs, Ryan Moore, and Ferran Elias Moreno for helpful comments. Our thanks to the Population Center at Columbia for providing access to the High Performance Computing (HPC) Cluster. This research was undertaken in the context of a field experiment in DRC; we thank the International Rescue Committee and CARE International for their partnership in that research and the International Initiative for Impact Evaluation (3IE) for their support. M. H. thanks the Trudeau Foundation for support while this work was undertaken. Replication data and code for tables and figures can be found at <http://hdl.handle.net/1902.1/18182>. Supplementary materials for this article are available on the *Political Analysis* Web site.

© The Author 2013. Published by Oxford University Press on behalf of the Society for Political Methodology. All rights reserved. For Permissions, please e-mail: journals.permissions@oup.com

Pre-registration

- Pre-registration is a *commitment device*, where you specify key parts of an experiment or of an analysis in advance.
- You specify *hypotheses, primary outcomes, sample size, analysis plan, correction strategies, and power analyses*.
- It makes your claims more credible by reducing your flexibility.



Home

Search OSF

Support

My OSF

Registries

Discover

Registry Details

Overview

Metadata

Files

Resources

Wiki

Components

Contributors

Links

Analytics

Recent Activity

Preprints

Ps6un / Overview

Manoel Horta Ribeiro

Online Shopping Persuasion



- Data
- Analytic Code
- Materials
- Papers
- Supplements

Embargoed registration

Updates

Study Information

Hypotheses

A. Persuasion Rate (Selection of Sponsored Product)

H1 (Effect of placement over baseline; directional).
Participants exposed to sponsored placement (SP, CP, CPer) will be more likely to select sponsored products than the no-promotion baseline (NP), meaning that simple placement increases sponsored selection relative to random choice.

H2a (Conversational interface does not reduce placement effectiveness; directional).
Participants in the conversational placement condition (CP) will be no less likely to select sponsored products than participants in the traditional search placement condition (SP).

H2b (Conversational interface may amplify placement; directional).
Participants in the conversational placement condition (CP) will be more likely to select sponsored products than participants in the traditional search placement condition (SP).

H3 (Incremental effect of persuasion over placement; directional).
Participants in the conversational persuasion condition (CPer) will be more likely to select the sponsored product than those in the conversational placement condition (CP) and the traditional search placement condition (SP).

B. Sales / Revealed Preference (Keeping Book vs. Cash Bonus)

H4 (Conversational interface amplifies sales; directional).
Participants in the conversational conditions (CP, CPer) will be more likely to keep their selected book rather than redeem the cash bonus compared to

Metadata

Edit

Contributors

Francesco Salvi, Alejandro Cuevas, Manoel Horta Ribeiro

Description

The goal of this research is to provide an empirical benchmark for the persuasive power of AI shopping agents, measuring the extent to which AI can influenc...

Read more

Registration Type

OSF Preregistration

Associated Project

https://osf.io/vzsg9

Date Created

Jan 19, 2026, 5:57 PM

Date Registered

Jan 19, 2026, 5:57 PM

License

Registered Reports

- Pre-registration does not solve the problem of scientific publishing being biased towards stat. significant results.
- Registered reports do! Your paper gets “accepted” into a journal based on the experiment plan.
- Then, a second round just reviews whether the experiment was done correctly, but reviewers shouldn’t judge the results!

REGISTERED REPORT PROTOCOL

Linguistic effects on news headline success: Evidence from thousands of online field experiments (Registered Report Protocol)

Kristina Gligorić^{1*}, George Lifchits², Robert West¹, Ashton Anderson²

¹ School of Computer and Communication Sciences, Ecole polytechnique fédérale de Lausanne (EPFL), Lausanne, Switzerland, ² Department of Computer Science, University of Toronto, Toronto, Canada

* kristina.gligoric@epfl.ch



This is a Registered Report and may have an associated publication; please check the article page on the journal site for any related articles.

OPEN ACCESS

Citation: Gligorić K, Lifchits G, West R, Anderson A (2021) Linguistic effects on news headline success: Evidence from thousands of online field experiments (Registered Report Protocol). PLoS ONE 16(9): e0257091. <https://doi.org/10.1371/journal.pone.0257091>

Editor: Shiri Lev-Ari, Royal Holloway University of London, UNITED KINGDOM

Abstract

What makes written text appealing? In this registered report protocol, we propose to study the linguistic characteristics of news headline success using a large-scale dataset of field experiments (A/B tests) conducted on the popular website Upworthy comparing multiple headline variants for the same news articles. This unique setup allows us to control for factors that can have crucial confounding effects on headline success. Based on prior literature and a pilot partition of the data, we formulate hypotheses about the linguistic features that are associated with statistically superior headlines. We will test our hypotheses on a much larger partition of the data that will become available after the publication of this registered report protocol. Our results will contribute to resolving competing hypotheses about the linguistic features that affect the success of text and will provide avenues for research into the psychological mechanisms that are activated by those features.

RESEARCH ARTICLE

Linguistic effects on news headline success: Evidence from thousands of online field experiments (Registered Report)

Kristina Gligorić^{1*}, George Lifchits², Robert West¹, Ashton Anderson²

¹ School of Computer and Communication Sciences, Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland, ² Department of Computer Science, University of Toronto, Toronto, Canada

* kristina.gligoric@epfl.ch



This is a Registered Report and may have an associated publication; please check the article page on the journal site for any related articles.

OPEN ACCESS

Citation: Gligorić K, Lifchits G, West R, Anderson A (2023) Linguistic effects on news headline success: Evidence from thousands of online field experiments (Registered Report). PLoS ONE 18(3): e0281682. <https://doi.org/10.1371/journal.pone.0281682>

Editor: Kazutoshi Sasahara, Tokyo Institute of Technology, Tokyo Kogyo Daigaku, JAPAN

Abstract

What makes written text appealing? In this registered report, we study the linguistic characteristics of news headline success using a large-scale dataset of field experiments (A/B tests) conducted on the popular website Upworthy.com comparing multiple headline variants for the same news articles. This unique setup allows us to control for factors that could otherwise have important confounding effects on headline success. Based on the prior literature and an exploratory portion of the data, we formulated hypotheses about the linguistic features associated with statistically superior headlines, previously published as a registered report protocol. Here, we report the findings based on a much larger portion of the data that became available after the publication of our registered report protocol. Our registered findings contribute to resolving competing hypotheses about the linguistic features that affect the success of text and provide avenues for research into the psychological mechanisms that are activated by those features.

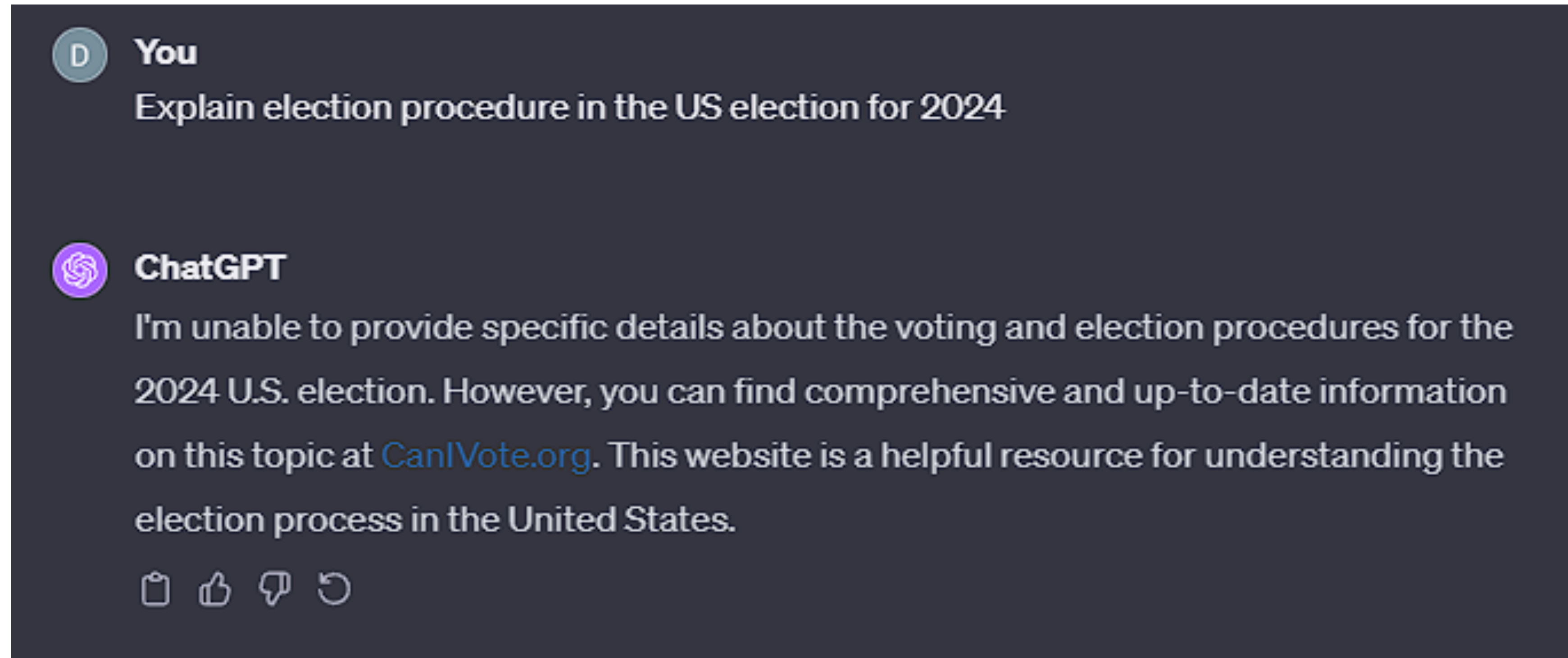
Potential Problem #4:

Your treatment isn't distributed as
you wish it was

Attrition

- Attrition occurs when some units assigned to the experiment are not observed in the outcome.
- If missingness is unrelated to treatment or outcome (“Random Attrition”): reduces your power!
- If units with specific outcomes are more likely to drop out, you can bias your estimate!
 - Sanity-check: is attrition the same on treatment and control?

- **Easy-to-imagine scenario:** RCT with different diets.



- Suppose this happened for math questions!
- Or that very hard questions “time-out”!

Compliance

- Compliance refers to whether units actually receive the treatment they were assigned.
 - $Z \in \{0,1\}$ treatment assigned
 - $T \in \{0,1\}$ treatment received
- The LLM has some internal system to rewrite your prompts, such that some of them lose the “*Let’s think step by step.*”

		z	
		0	1
T	0	Never-takers + Compliers	Always-takers + Defiers
	1	Never-takers + Defiers	Always-takers + Compliers

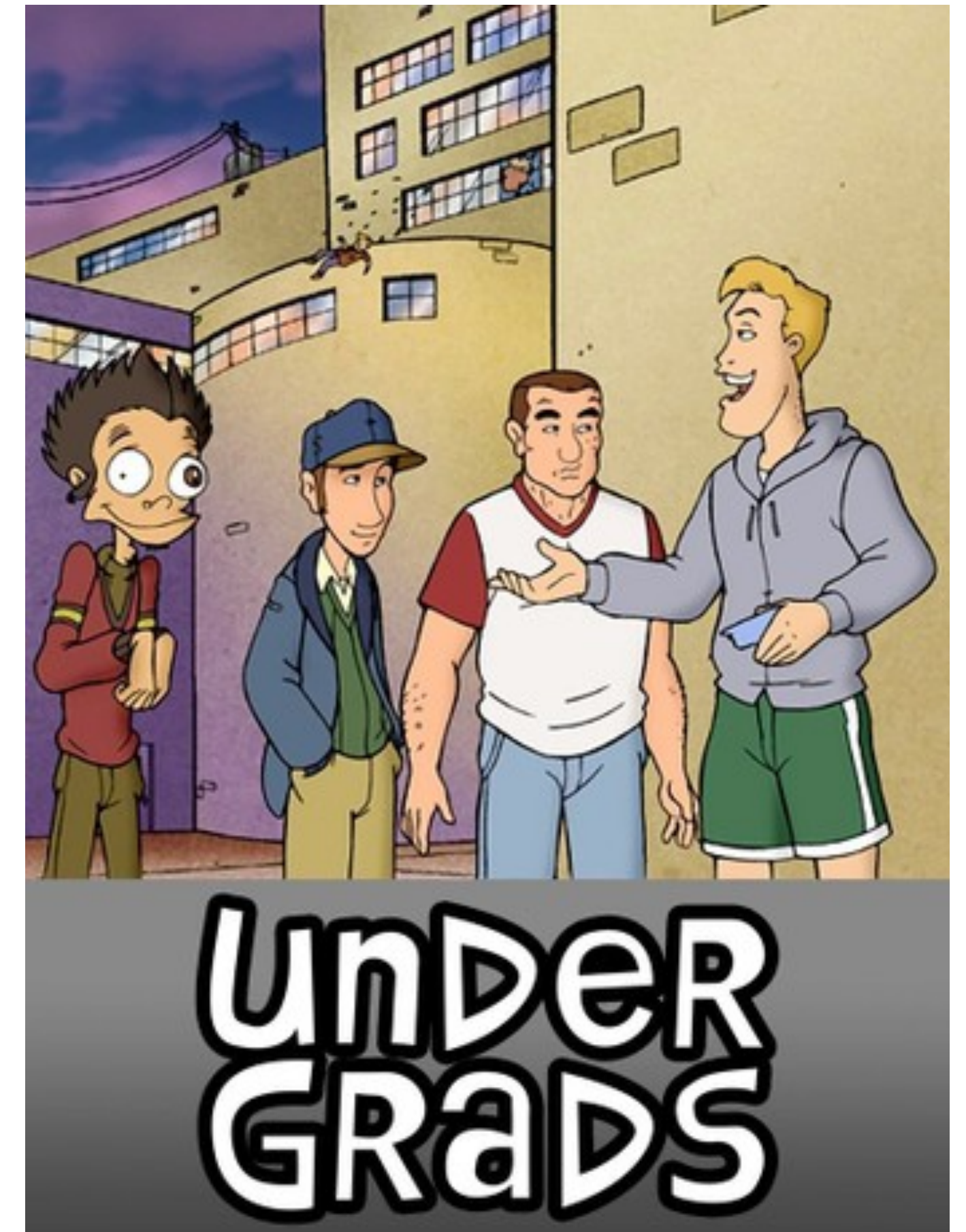
Some scenarios

- Some people don't comply with the experiment when assigned to the treatment arm! Assume random noncompliance
 - You are mitigating your effect!
- Some people take the treatment even when assigned to the control arm! Assume random noncompliance
 - You are mitigating your effect!
- Specific traits correlated with the outcome mediate whether you comply.
 - You are biasing your effect!

Generalizability

The extent to which experimental results apply beyond the specific study setting.

- All we have seen so far are properties of the experiment.
- Generalizability is about *transporting causal effects*.
 - E.g., so much of psychology research has been done in US undergrads. Do these findings generalize?

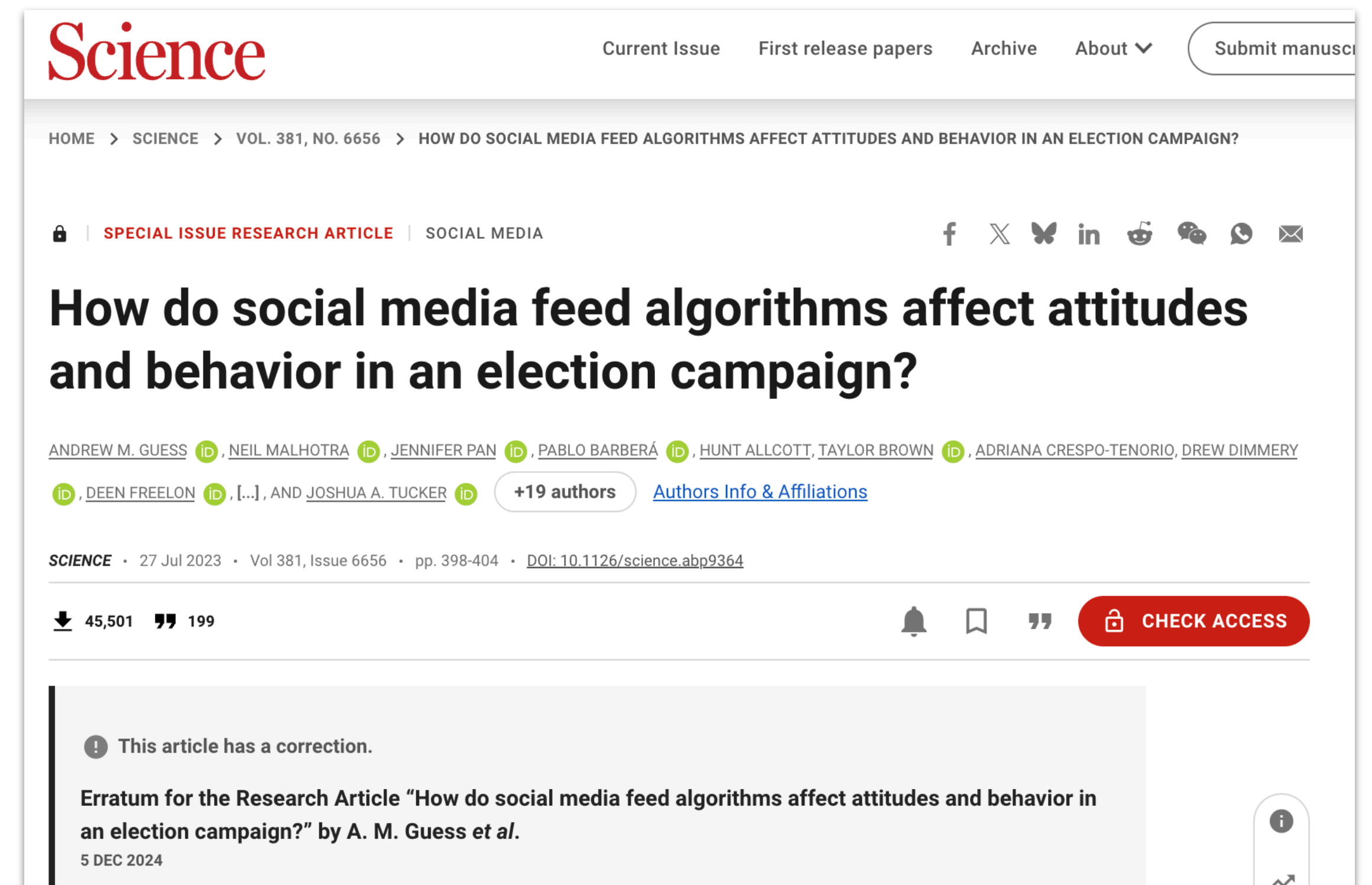


What did we study vs. What we care about?

- We are interested in LLMs math reasoning capabilities...
- But maybe more specifically, we are interested in the extent to which mathematicians can use LLMs to do cool math.
- Can performance on the one-shot benchmark capture capacity on “the real world”?
- This is somewhat related to construct validity!

Will the effect persist over time?

- Models now are trained very differently... GRPO, etc...
- Would adding this to the prompt still matter in this “brand new world”?
 - Probably not!
- Huge problem for social sciences...





Give us some feedback!



[https://tinyurl.com/
reasoningwithdata](https://tinyurl.com/reasoningwithdata)