Experiments

Manoel Horta Ribeiro
manoel@cs.princeton.edu

humans & machines lab

COS 598D / Spring 2026

Recap: Experiments Rock

- We show we can estimate the ATE:

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

- No exchangeability:

$$(Y_{ir}^0, Y_{ir}^1) \not\perp T_{ir}\,.$$

- Manoel's thoughts:
  - In the age of AI agents, we are now in the small *n* regimen.
  - Variability is a huge issue (not only bias)
  - I really liked their recommendations, some unintuitive ones for me were:
    - Software/hardware stack
    - Answer extraction

A Sober Look at Progress in Language Model Reasoning: Pitfalls and Paths to Reproducibility

Andreas Hochlehnert[1*] Hardik Bhatnagar[1*] Vishaal Udandarao[1,2°]
Samuel Albanie Ameya Prabhu[1†] Matthias Bethge[1†]

[1]Tübingen AI Center, University of Tübingen [2] University of Cambridge

Leaderboard Code Eval Logs

Abstract

Reasoning has emerged as the next major frontier for language models (LMs), with rapid advances from both academic and industrial labs. However, this progress often outpaces methodological rigor, with many evaluations relying on benchmarking practices that lack transparency, robustness, or statistical grounding. In this work, we conduct a comprehensive empirical study and find that current mathematical reasoning benchmarks are highly sensitive to subtle implementation choices—including decoding parameters, random seeds, prompt formatting, and even hardware and software configurations. Performance gains reported in recent studies frequently hinge on unclear comparisons or unreported sources of variance. To address these issues, we propose a standardized evaluation framework with clearly defined best practices and reporting standards. Using this framework, we reassess recent methods and find that most reinforcement learning (RL) approaches yield only modest improvements—far below prior claims—and are prone to overfitting, especially on small-scale benchmarks like AIME'24. In contrast, supervised finetuning (SFT) methods show consistently stronger generalization in the settings we study. To foster reproducibility, we release all code, prompts, and model outputs, for reasoning benchmarks, establishing more rigorous foundations for future work.

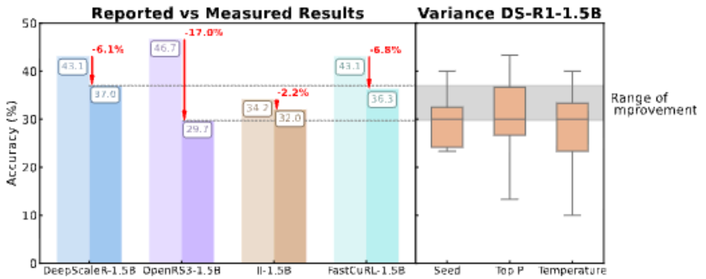


Figure 1: **The Sombre State of LM Reasoning for Math.** *(left)* when re-evaluating recent 1.5B reasoning-enhanced models on AIME-24 using a standardized framework (see Section 4), we find substantial drops to reported results in the original papers, *(right)* the observed improvements from recent methods (gray highlighted area) fall entirely within the variance range (orange box plots) of DeepSeek-R1 1.5B model performance. This suggests that these methods do not significantly outperform the base model—underscoring the importance of rigorous, multi-seed evaluation protocols for obtaining reliable performance estimates.

*equal contribution, ° core contributor, †equal advising

1

Not the first crisis in sight

Open access, freely available online

Essay

Why Most Published Research Findings Are False

John P. A. Ioannidis

Summary

There is increasing concern that most current published research findings are false. The probability that a research claim is true may depend on study power and bias, the number of other studies on the same question, and, importantly, the ratio of true to no relationships among the relationships probed in each scientific field. In this framework, a research finding is less likely to be true when the studies conducted in a field are smaller; when effect sizes are smaller; when there is a greater number and lesser preselection of tested relationships; where there is greater flexibility in designs, definitions, outcomes, and analytical modes; when there is greater financial and other interest and prejudice; and when more teams are involved in a scientific field in chase of statistical significance. Simulations show that for most study designs and settings, it is more likely for a research claim to be false than true. Moreover, for many current scientific fields, claimed research findings may often be simply accurate measures of the prevailing bias. In this essay, I discuss the implications of these problems for the conduct and interpretation of research.

Published research findings are sometimes refuted by subsequent evidence, with ensuing confusion and disappointment. Refutation and controversy is seen across the range of research designs, from clinical trials and traditional epidemiological studies [1–3] to the most modern molecular research [4,5]. There is increasing concern that in modern research, false findings may be the majority or even the vast majority of published research claims [6–8]. However, this should not be surprising. It can be proven that most claimed research findings are false. Here I will examine the key factors that influence this problem and some corollaries thereof.

The Essay section contains opinion pieces on topics of broad interest to a general medical audience.

Modeling the Framework for False Positive Findings

Several methodologists have pointed out [9–11] that the high rate of nonreplication (lack of confirmation) of research discoveries is a consequence of the convenient, yet ill-founded strategy of claiming conclusive research findings solely on the basis of a single study assessed by formal statistical significance, typically for a *p*-value less than 0.05. Research is not most appropriately represented and summarized by *p*-values, but, unfortunately, there is a widespread notion that medical research articles

It can be proven that most claimed research findings are false.

should be interpreted based only on *p*-values. Research findings are defined here as any relationship reaching formal statistical significance, e.g., effective interventions, informative predictors, risk factors, or associations. "Negative" research is also very useful. "Negative" is actually a misnomer, and the misinterpretation is widespread. However, here we will target relationships that investigators claim exist, rather than null findings.

As has been shown previously, the probability that a research finding is indeed true depends on the prior probability of it being true (before doing the study), the statistical power of the study, and the level of statistical significance [10,11]. Consider a 2 × 2 table in which research findings are compared against the gold standard of true relationships in a scientific field. In a research field both true and false hypotheses can be made about the presence of relationships. Let R be the ratio of the number of "true relationships" to "no relationships" among those tested in the field. R is characteristic of the field and can vary a lot depending on whether the field targets highly likely relationships or searches for only one or a few true relationships among thousands and millions of hypotheses that may be postulated. Let us also consider, for computational simplicity, circumscribed fields where either there is only one true relationship (among many that can be hypothesized) or the power is similar to find any of the several existing true relationships. The pre-study probability of a relationship being true is $R/(R + 1)$. The probability of a study finding a true relationship reflects the power $1 - \beta$ (one minus the Type II error rate). The probability of claiming a relationship when none truly exists reflects the Type I error rate, α. Assuming that c relationships are being probed in the field, the expected values of the 2 × 2 table are given in Table 1. After a research finding has been claimed based on achieving formal statistical significance, the post-study probability that it is true is the positive predictive value, PPV. The PPV is also the complementary probability of what Wacholder et al. have called the false positive report probability [10]. According to the 2 × 2 table, one gets PPV = $(1 - \beta)R/(R - \beta R + \alpha)$. A research finding is thus

Citation: Ioannidis JPA (2005) Why most published research findings are false. PLoS Med 2(8): e124.

Abbreviation: PPV, positive predictive value

John P.A. Ioannidis is in the Department of Hygiene and Epidemiology, University of Ioannina School of Medicine, Ioannina, Greece, and Institute for Clinical Research and Health Policy Studies, Department of Medicine, Tufts-New England Medical Center, Tufts University School of Medicine, Boston, Massachusetts, United States of America. E-mail: jioannid@cc.uoi.gr

Competing Interests: The author has declared that no competing interests exist.

DOI: 10.1371/journal.pmed.0020124

SCIENCE

Psychology's Replication Crisis Is Running Out of Excuses

Another big project has found that only half of studies can be repeated. And this time, the usual explanations fall flat.

By Ed Yong

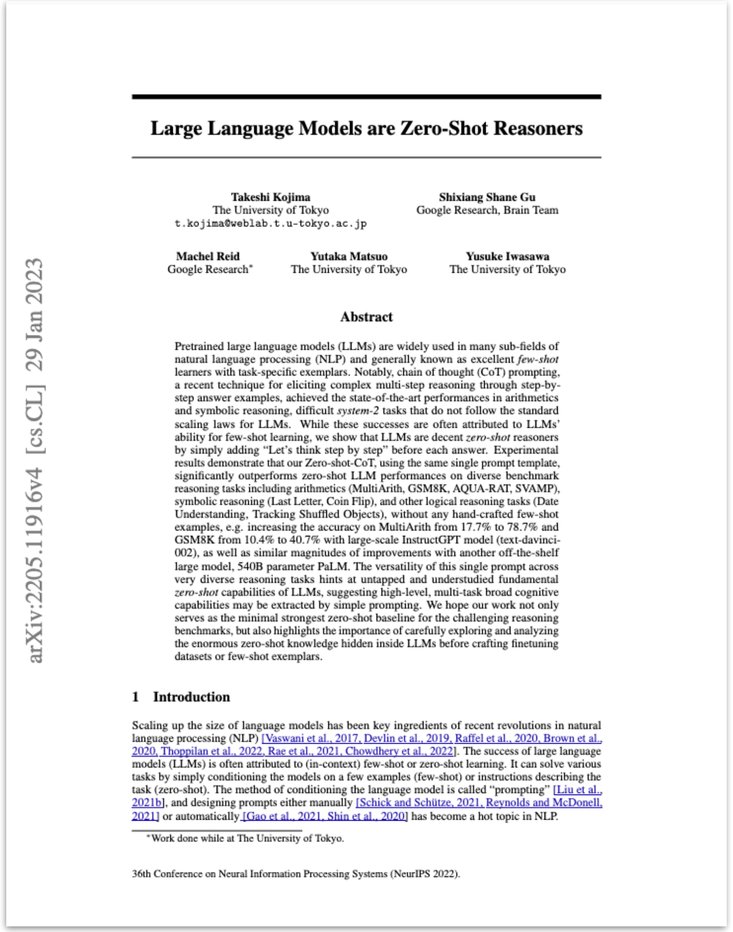
The Thinker, by Auguste Rodin (Jason Lee / Reuters)

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.**



arXiv:2205.11916v4 [cs.CL] 29 Jan 2023

Large Language Models are Zero-Shot Reasoners

Takeshi Kojima
The University of Tokyo
t.kojima@weblab.t.u-tokyo.ac.jp

Shixiang Shane Gu
Google Research, Brain Team

Machel Reid
Google Research*

Yutaka Matsuo
The University of Tokyo

Yusuke Iwasawa
The University of Tokyo

Abstract

Pretrained large language models (LLMs) are widely used in many sub-fields of natural language processing (NLP) and generally known as excellent *few-shot* learners with task-specific exemplars. Notably, chain of thought (CoT) prompting, a recent technique for eliciting complex multi-step reasoning through step-by-step answer examples, achieved the state-of-the-art performances in arithmetics and symbolic reasoning, difficult *system-2* tasks that do not follow the standard scaling laws for LLMs. While these successes are often attributed to LLMs' ability for few-shot learning, we show that LLMs are decent *zero-shot* reasoners by simply adding "Let's think step by step" before each answer. Experimental results demonstrate that our Zero-shot-CoT, using the same single prompt template, significantly outperforms zero-shot LLM performances on diverse benchmark reasoning tasks including arithmetics (MultiArith, GSM8K, AQUA-RAT, SVAMP), symbolic reasoning (Last Letter, Coin Flip), and other logical reasoning tasks (Date Understanding, Tracking Shuffled Objects), without any hand-crafted few-shot examples, e.g. increasing the accuracy on MultiArith from 17.7% to 78.7% and GSM8K from 10.4% to 40.7% with large-scale InstructGPT model (text-davinci-002), as well as similar magnitudes of improvements with another off-the-shelf large model, 540B parameter PaLM. The versatility of this single prompt across very diverse reasoning tasks hints at untapped and understudied fundamental *zero-shot* capabilities of LLMs, suggesting high-level, multi-task broad cognitive capabilities may be extracted by simple prompting. We hope our work not only serves as the minimal strongest zero-shot baseline for the challenging reasoning benchmarks, but also highlights the importance of carefully exploring and analyzing the enormous zero-shot knowledge hidden inside LLMs before crafting finetuning datasets or few-shot exemplars.

1 Introduction

Scaling up the size of language models has been key ingredients of recent revolutions in natural language processing (NLP) [Vaswani et al., 2017, Devlin et al., 2019, Raffel et al., 2020, Brown et al., 2020, Thoppilan et al., 2022, Rae et al., 2021, Chowdhery et al., 2022]. The success of large language models (LLMs) is often attributed to (in-context) few-shot or zero-shot learning. It can solve various tasks by simply conditioning the models on a few examples (few-shot) or instructions describing the task (zero-shot). The method of conditioning the language model is called "prompting" [Liu et al., 2021b], and designing prompts either manually [Schick and Schütze, 2021, Reynolds and McDonell, 2021] or automatically [Gao et al., 2021, Shin et al., 2020] has become a hot topic in NLP.

*Work done while at The University of Tokyo.

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

Potential Problem #1:

You measure the wrong thing

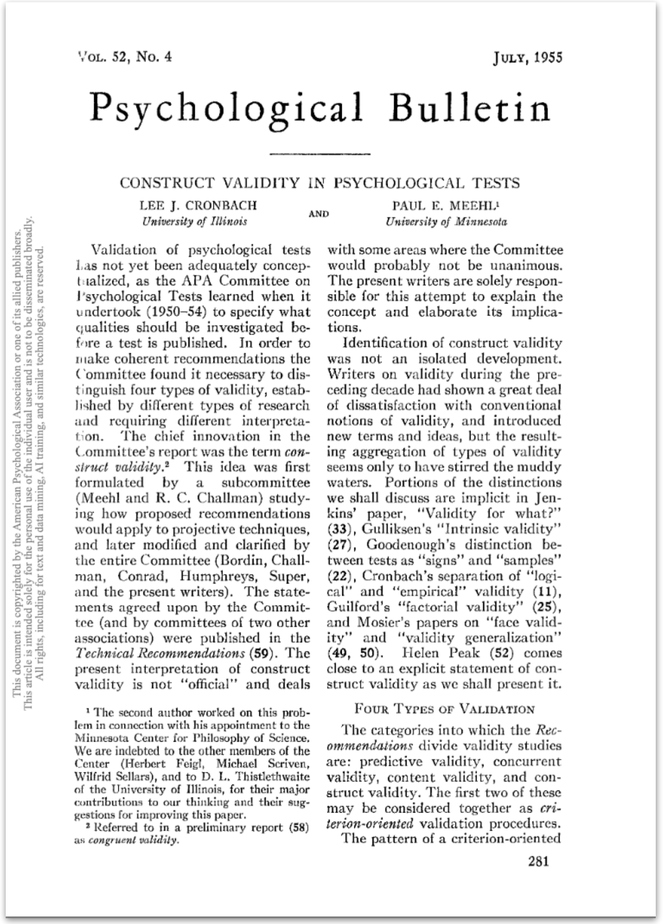
Construct vs. Measurement

- 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.



VOL. 52, No. 4 JULY, 1955

Psychological Bulletin

CONSTRUCT VALIDITY IN PSYCHOLOGICAL TESTS

LEE J. CRONBACH
University of Illinois
AND
PAUL E. MEEHL[1]
University of Minnesota

Validation of psychological tests has not yet been adequately conceptualized, as the APA Committee on Psychological Tests learned when it undertook (1950–54) to specify what qualities should be investigated before a test is published. In order to make coherent recommendations the Committee found it necessary to distinguish four types of validity, established by different types of research and requiring different interpretation. The chief innovation in the Committee's report was the term *construct validity.*[2] This idea was first formulated by a subcommittee (Meehl and R. C. Challman) studying how proposed recommendations would apply to projective techniques, and later modified and clarified by the entire Committee (Bordin, Challman, Conrad, Humphreys, Super, and the present writers). The statements agreed upon by the Committee (and by committees of two other associations) were published in the *Technical Recommendations* (59). The present interpretation of construct validity is not "official" and deals with some areas where the Committee would probably not be unanimous. The present writers are solely responsible for this attempt to explain the concept and elaborate its implications.

Identification of construct validity was not an isolated development. Writers on validity during the preceding decade had shown a great deal of dissatisfaction with conventional notions of validity, and introduced new terms and ideas, but the resulting aggregation of types of validity seems only to have stirred the muddy waters. Portions of the distinctions we shall discuss are implicit in Jenkins' paper, "Validity for what?" (33), Gulliksen's "Intrinsic validity" (27), Goodenough's distinction between tests as "signs" and "samples" (22), Cronbach's separation of "logical" and "empirical" validity (11), Guilford's "factorial validity" (25), and Mosier's papers on "face validity" and "validity generalization" (49, 50). Helen Peak (52) comes close to an explicit statement of construct validity as we shall present it.

FOUR TYPES OF VALIDATION

The categories into which the *Recommendations* divide validity studies are: predictive validity, concurrent validity, content validity, and construct validity. The first two of these may be considered together as *criterion-oriented* validation procedures. The pattern of a criterion-oriented

[1] The second author worked on this problem in connection with his appointment to the Minnesota Center for Philosophy of Science. We are indebted to the other members of the Center (Herbert Feigl, Michael Scriven, Wilfrid Sellars), and to D. L. Thistlethwaite of the University of Illinois, for their major contributions to our thinking and their suggestions for improving this paper.

[2] Referred to in a preliminary report (58) as *congruent validity.*

Cronbach & Meehl (1955)

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?

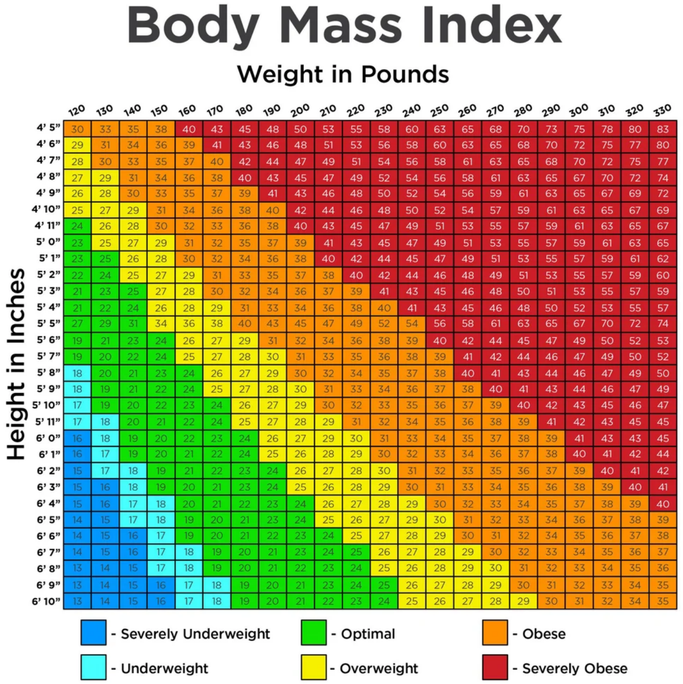
- Is the measurement *driven by a different construct*?
  - Maybe the model isn't reasoning at all, it is just that when there are more tokens in the context, the model is more likely to give a correct answer?

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

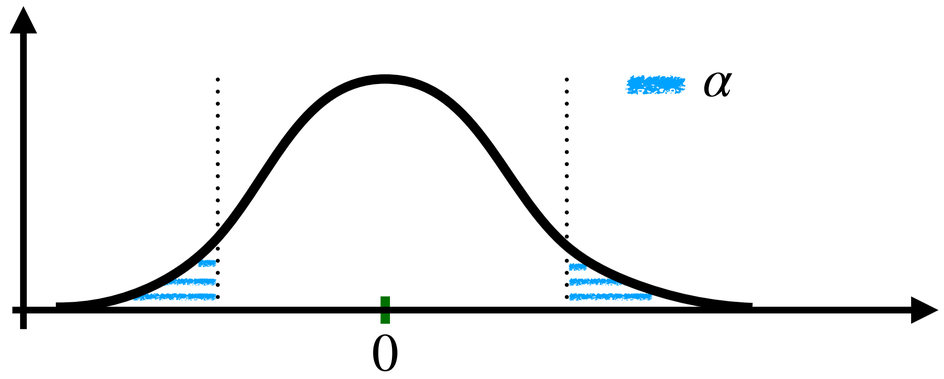
- But what if H_0 it is false and H_1 is true?
- Not all experiments can reject the null very often!
- Let β be the probability of "missing out" on an effect.
  - $1 - \beta$ = the probability we reject the null when it is false.
  - We call $1 - \beta$ the statistical power of an experiment.

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.

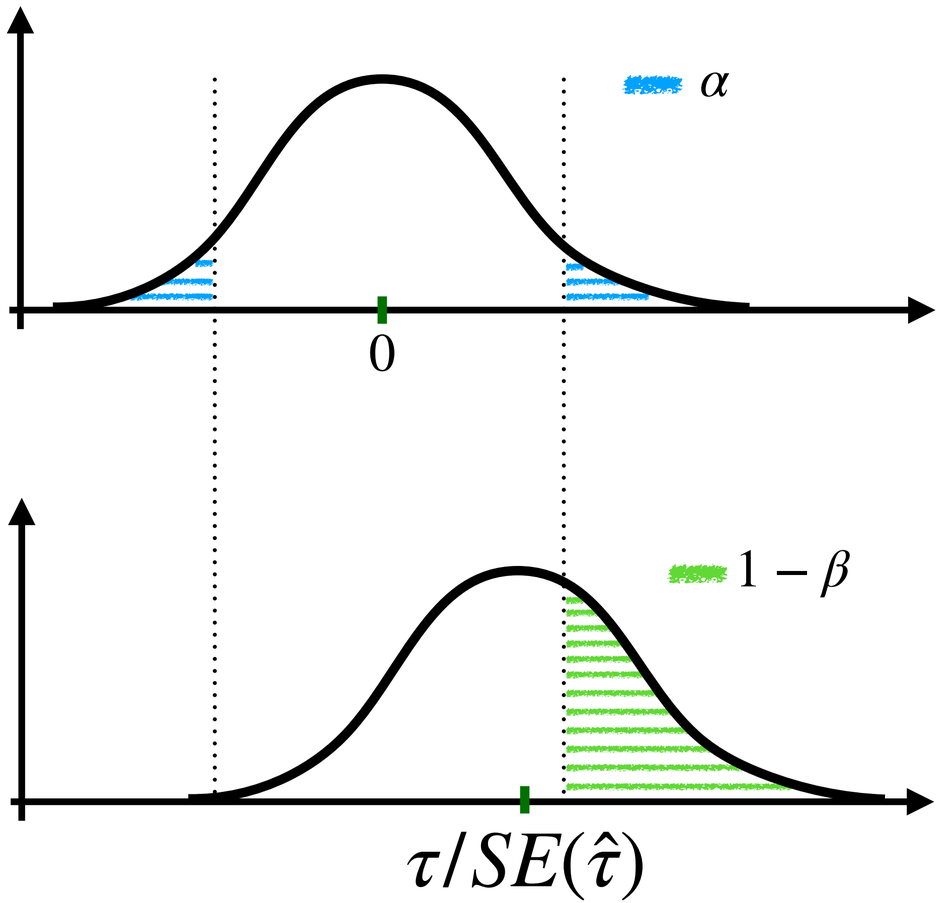


Power

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

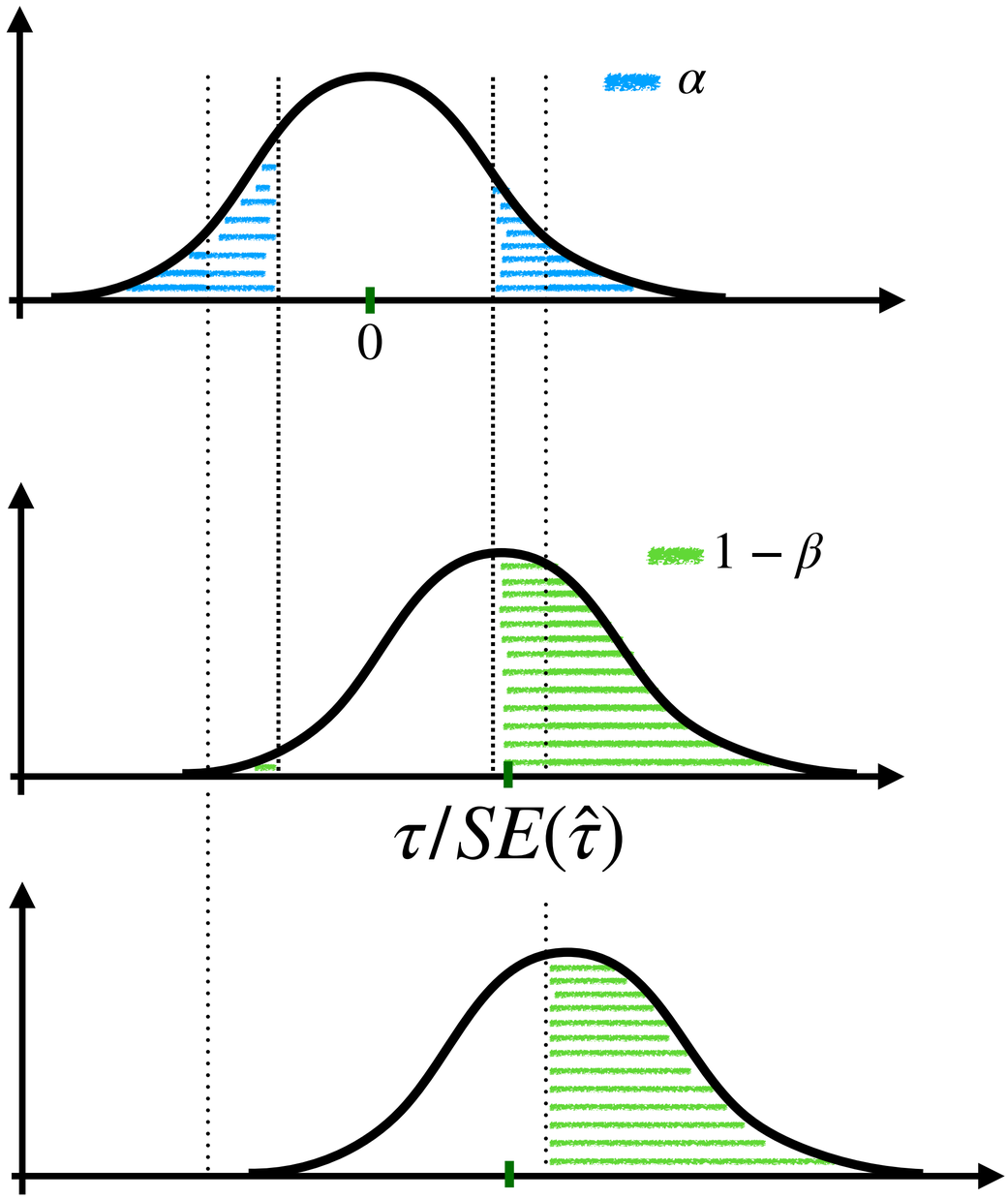
$$Z = \frac{\hat{\tau} - \tau}{SE(\hat{\tau})} = \frac{\hat{\tau} - 0}{SE(\hat{\tau})} \sim \mathcal{N}\left(\frac{\tau}{SE(\hat{\tau})}, 1\right)$$

- **Power**: orange area / total area



Increasing power

- Power depends on:
  - $Var(\hat{\tau})$, because $SE(\hat{\tau}) \sim Var(\hat{\tau})$
  - 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,

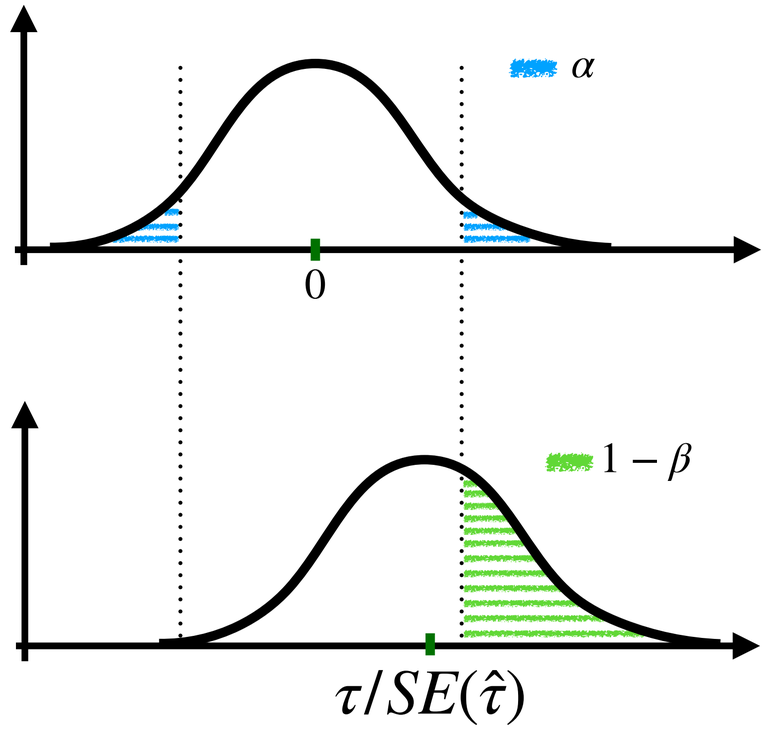
$$Z \sim \mathcal{N}\left(\tau/SE(\hat{\tau}),1\right)$$

Then, we have that power equals to:

$$1-\beta =$$

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

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



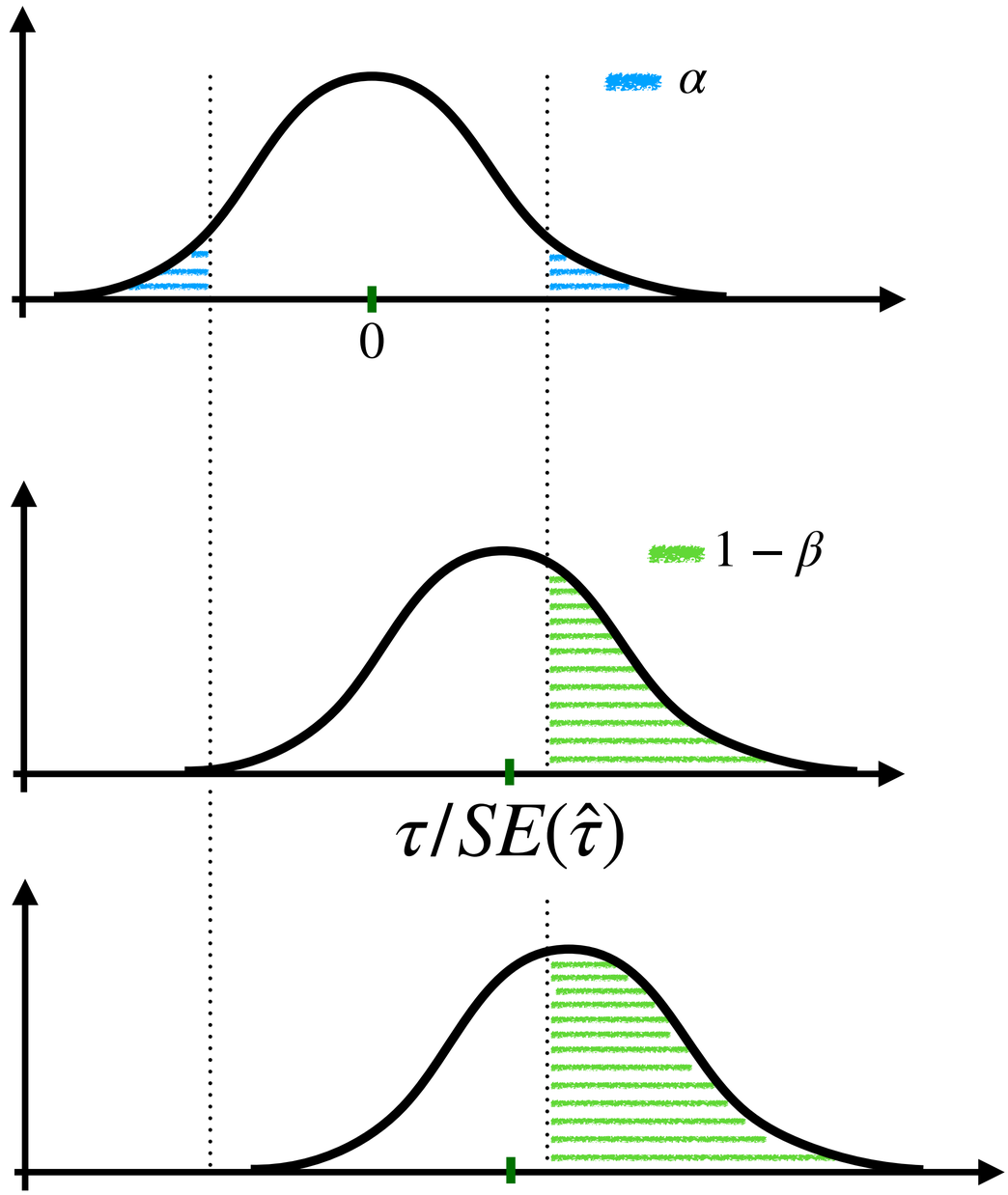
Power simulations

- 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"

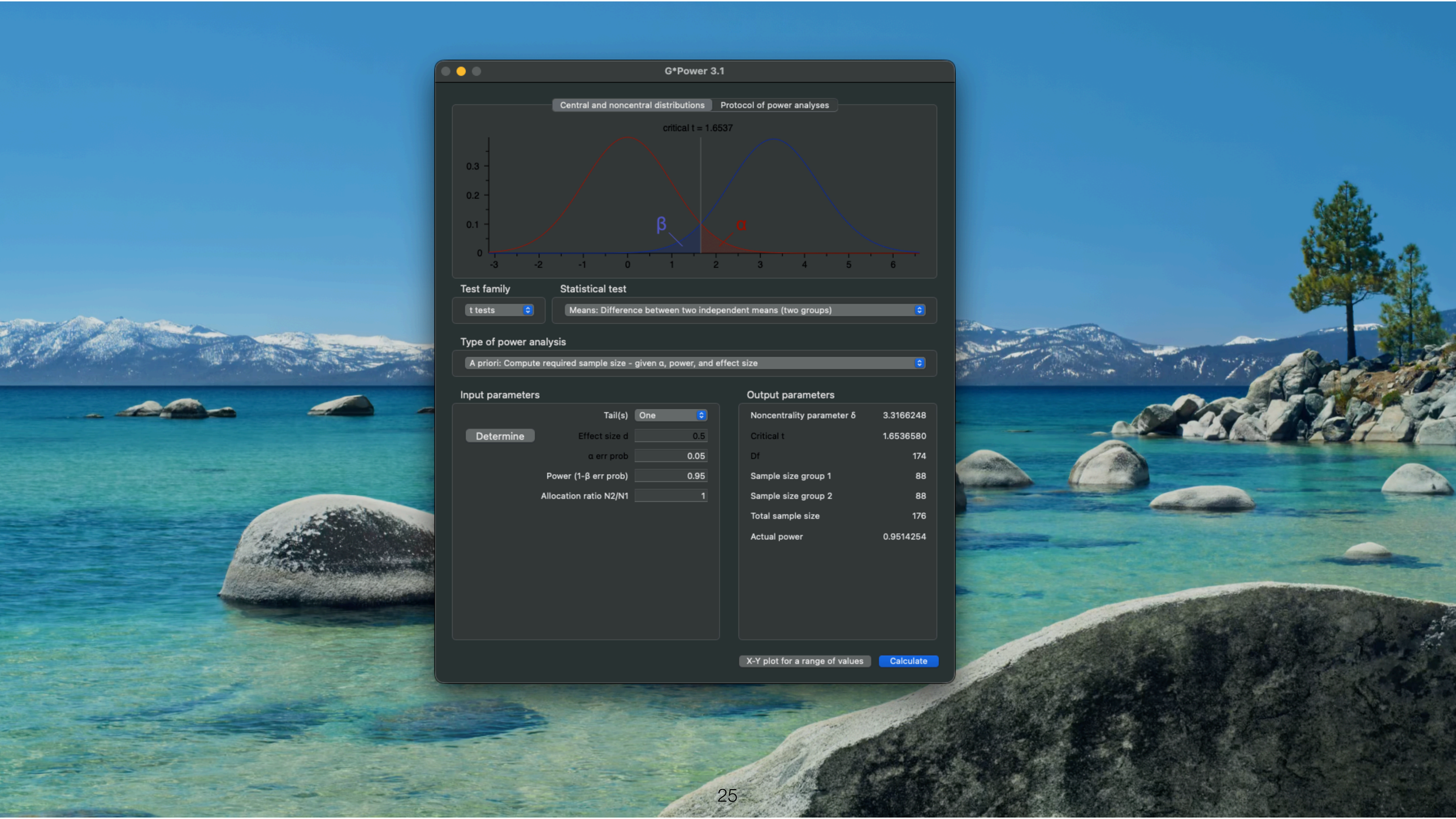
$$MDE = \left(z_{1-\alpha/2} + z_{1-\beta}\right)\frac{\sigma}{\sqrt{n}}$$



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.


G*Power 3.1
Central and noncentral distributions
Protocol of power analyses
critical t = 1.6537
0.3
0.2
0.1
0
β
α
-3
-2
-1
0
1
2
3
4
5
6
Test family
t tests
Statistical test
Means: Difference between two independent means (two groups)
Type of power analysis
A priori: Compute required sample size - given α, power, and effect size
Input parameters
Tail(s)
One
Determine
Effect size d
0.5
α err prob
0.05
Power (1-β err prob)
0.95
Allocation ratio N2/N1
1
Output parameters
Noncentrality parameter δ
3.3166248
Critical t
1.6536580
Df
174
Sample size group 1
88
Sample size group 2
88
Total sample size
176
Actual power
0.9514254
X-Y plot for a range of values
Calculate

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.

Open access, freely available online

Essay

Why Most Published Research Findings Are False

John P. A. Ioannidis

Summary

There is increasing concern that most current published research findings are false. The probability that a research claim is true may depend on study power and bias, the number of other studies on the same question, and, importantly, the ratio of true to no relationships among the relationships probed in each scientific field. In this framework, a research finding is less likely to be true when the studies conducted in a field are smaller; when effect sizes are smaller; when there is a greater number and lesser preselection of tested relationships; where there is greater flexibility in designs, definitions, outcomes, and analytical modes; when there is greater financial and other interest and prejudice; and when more teams are involved in a scientific field in chase of statistical significance. Simulations show that for most study designs and settings, it is more likely for a research claim to be false than true. Moreover, for many current scientific fields, claimed research findings may often be simply accurate measures of the prevailing bias. In this essay, I discuss the implications of these problems for the conduct and interpretation of research.

Published research findings are sometimes refuted by subsequent evidence, with ensuing confusion and disappointment. Refutation and controversy is seen across the range of research designs, from clinical trials and traditional epidemiological studies [1–3] to the most modern molecular research [4,5]. There is increasing concern that in modern research, false findings may be the majority or even the vast majority of published research claims [6–8]. However, this should not be surprising. It can be proven that most claimed research findings are false. Here I will examine the key factors that influence this problem and some corollaries thereof.

The Essay section contains opinion pieces on topics of broad interest to a general medical audience.

Modeling the Framework for False Positive Findings

Several methodologists have pointed out [9–11] that the high rate of nonreplication (lack of confirmation) of research discoveries is a consequence of the convenient, yet ill-founded strategy of claiming conclusive research findings solely on the basis of a single study assessed by formal statistical significance, typically for a p-value less than 0.05. Research is not most appropriately represented and summarized by p-values, but, unfortunately, there is a widespread notion that medical research articles

It can be proven that most claimed research findings are false.

should be interpreted based only on p-values. Research findings are defined here as any relationship reaching formal statistical significance, e.g., effective interventions, informative predictors, risk factors, or associations. "Negative" research is also very useful. "Negative" is actually a misnomer, and the misinterpretation is widespread. However, here we will target relationships that investigators claim exist, rather than null findings.

As has been shown previously, the probability that a research finding is indeed true depends on the prior probability of it being true (before doing the study), the statistical power of the study, and the level of statistical significance [10,11]. Consider a 2×2 table in which research findings are compared against the gold standard of true relationships in a scientific field. In a research field both true and false hypotheses can be made about the presence of relationships. Let R be the ratio of the number of "true relationships" to "no relationships" among those tested in the field. R is characteristic of the field and can vary a lot depending on whether the field targets highly likely relationships or searches for only one or a few true relationships among thousands and millions of hypotheses that may be postulated. Let us also consider, for computational simplicity, circumscribed fields where either there is only one true relationship (among many that can be hypothesized) or the power is similar to find any of the several existing true relationships. The pre-study probability of a relationship being true is $R/(R + 1)$. The probability of a study finding a true relationship reflects the power $1 - \beta$ (one minus the Type II error rate). The probability of claiming a relationship when none truly exists reflects the Type I error rate, α. Assuming that c relationships are being probed in the field, the expected values of the 2×2 table are given in Table 1. After a research finding has been claimed based on achieving formal statistical significance, the post-study probability that it is true is the positive predictive value, PPV. The PPV is also the complementary probability of what Wacholder et al. have called the false positive report probability [10]. According to the 2×2 table, one gets $\text{PPV} = (1 - \beta)R/(R - \beta R + \alpha)$. A research finding is thus

Citation: Ioannidis JPA (2005) Why most published research findings are false. PLoS Med 2(8): e124.

Abbreviation: PPV, positive predictive value

John P. A. Ioannidis is in the Department of Hygiene and Epidemiology, University of Ioannina School of Medicine, Ioannina, Greece, and Institute for Clinical Research and Health Policy Studies, Department of Medicine, Tufts-New England Medical Center, Tufts University School of Medicine, Boston, Massachusetts, United States of America. E-mail: jioannid@cc.uoi.gr

Competing Interests: The author has declared that no competing interests exist.

DOI: 10.1371/journal.pmed.0020124

Correcting for multiple comparisons

- Let $H_1, \ldots, H_m$ be a family of null hypotheses
- Let $\mathcal{I}_0$ be the set of indices where the null is actually true.
- Let $p_1, \ldots, p_m$ 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 $p_i \leq \alpha/m$

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

- 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 ith 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."

Political Analysis (2013) 21:1–20
doi:10.1093/pan/mps021

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.

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.

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

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!

PLOS ONE

REGISTERED REPORT PROTOCOL

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

Kristina Gligorić[1]*, George Lifchits[2], Robert West[1], Ashton Anderson[2]

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

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.

PLOS ONE

RESEARCH ARTICLE

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

Kristina Gligorić[1]*, George Lifchits[2], Robert West[1], Ashton Anderson[2]

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

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

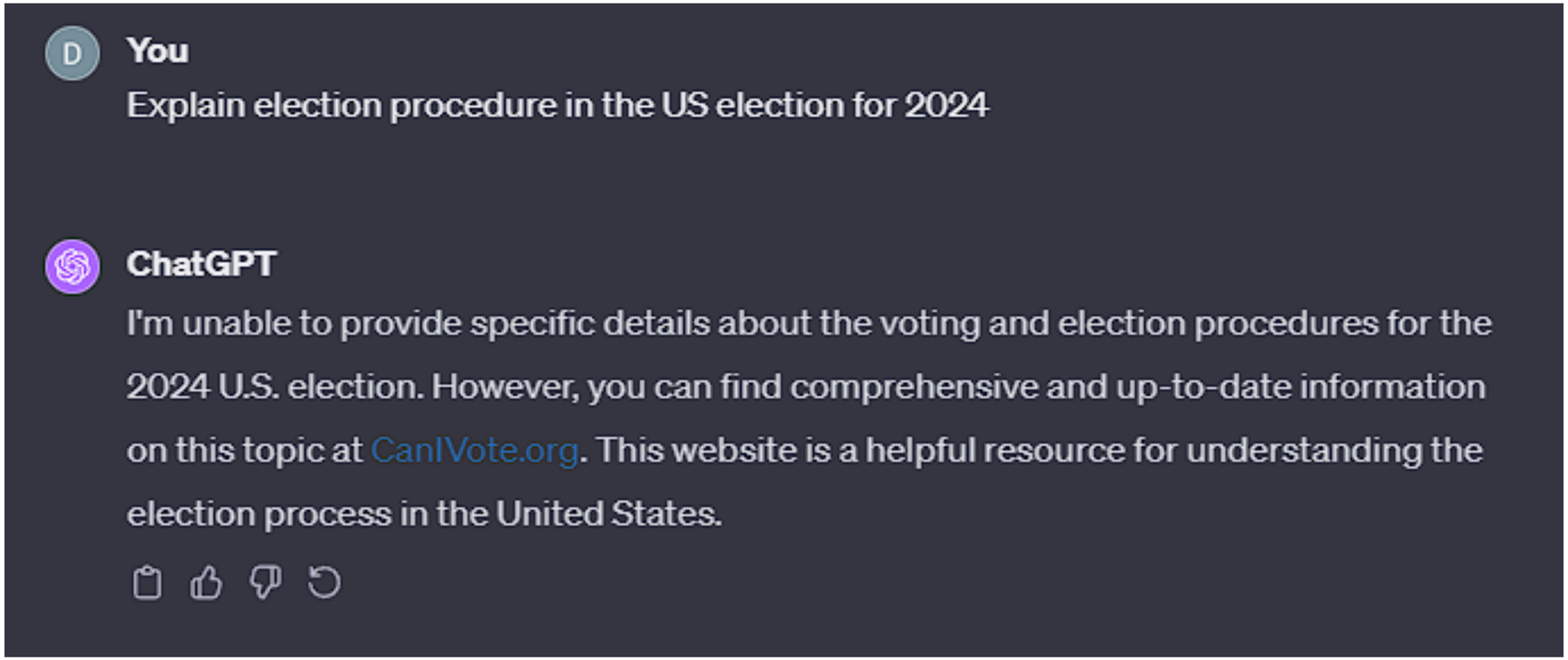
What makes written text appealing? In this registered report, we study the linguistic characteristics of news headline success using a large-scale dataset of field experiments (A/B tests) conducted on the popular website Upworthy.com comparing multiple headline variants for the same news articles. This 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!
  - You are mitigating your effect!
- Some people take the treatment even when assigned to the control arm!
  - You are mitigating your effect!
- Specific traits correlated with the outcome mediate whether you comply.
  - You are biasing your effect!