PODS Lab 11: Replication Crisis, Effect Size, and Power

Hamza Alshamy

Center for Data Science, NYU
ha2486@nyu.edu

04/11/25

1. Replication Crisis (P-hacking)
2. Effect Size (Cohen's d)
3. Power

Incentives... and the replication crisis

Incentives... and the replication crisis

- **Incentives:** Anything that motivates individuals, businesses, or governments to take specific actions or make decisions, typically based on the potential for gaining a benefit or avoiding a cost.

Incentives... and the replication crisis

- **Incentives:** Anything that motivates individuals, businesses, or governments to take specific actions or make decisions, typically based on the potential for gaining a benefit or avoiding a cost.
- **Question:** What is your incentive to go to class?

Incentives... and the replication crisis

- **Incentives:** Anything that motivates individuals, businesses, or governments to take specific actions or make decisions, typically based on the potential for gaining a benefit or avoiding a cost.
- **Question:** What is your incentive to go to class?
- **Question:** What role do incentives have in academia and publishing?

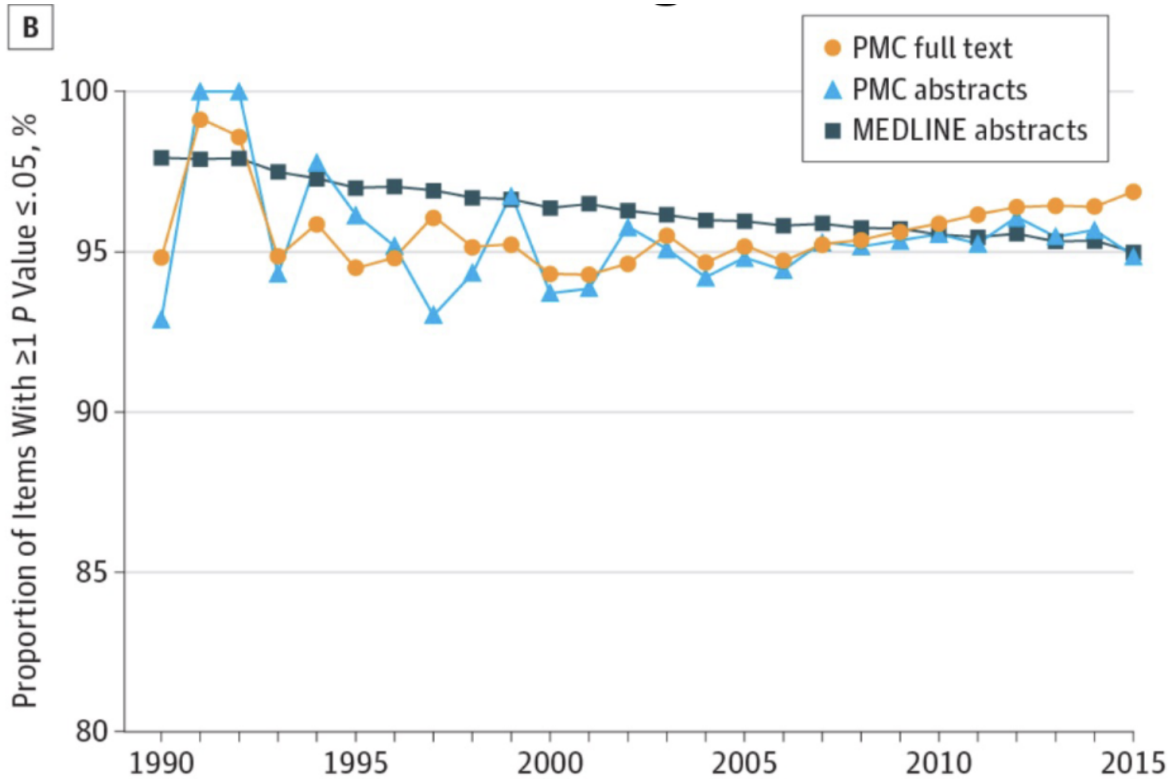
Incentives... and the replication crisis

- **Incentives:** Anything that motivates individuals, businesses, or governments to take specific actions or make decisions, typically based on the potential for gaining a benefit or avoiding a cost.
- **Question:** What is your incentive to go to class?
- **Question:** What role do incentives have in academia and publishing?

- **How is the replication crisis possible?**
  - Researchers are often pressured to produce significant results in order to publish papers in high-impact journals, secure jobs, or obtain grants and funding for future projects.
  - This **incentive system** leads to a focus on statistical significance rather than robust, replicable findings
  - **Result:** Researchers may resort to questionable practices to achieve publishable results.

P-value as marker for significance

Proportion of scientific articles that reported at least one statistically significant result ($P < 0.05$)



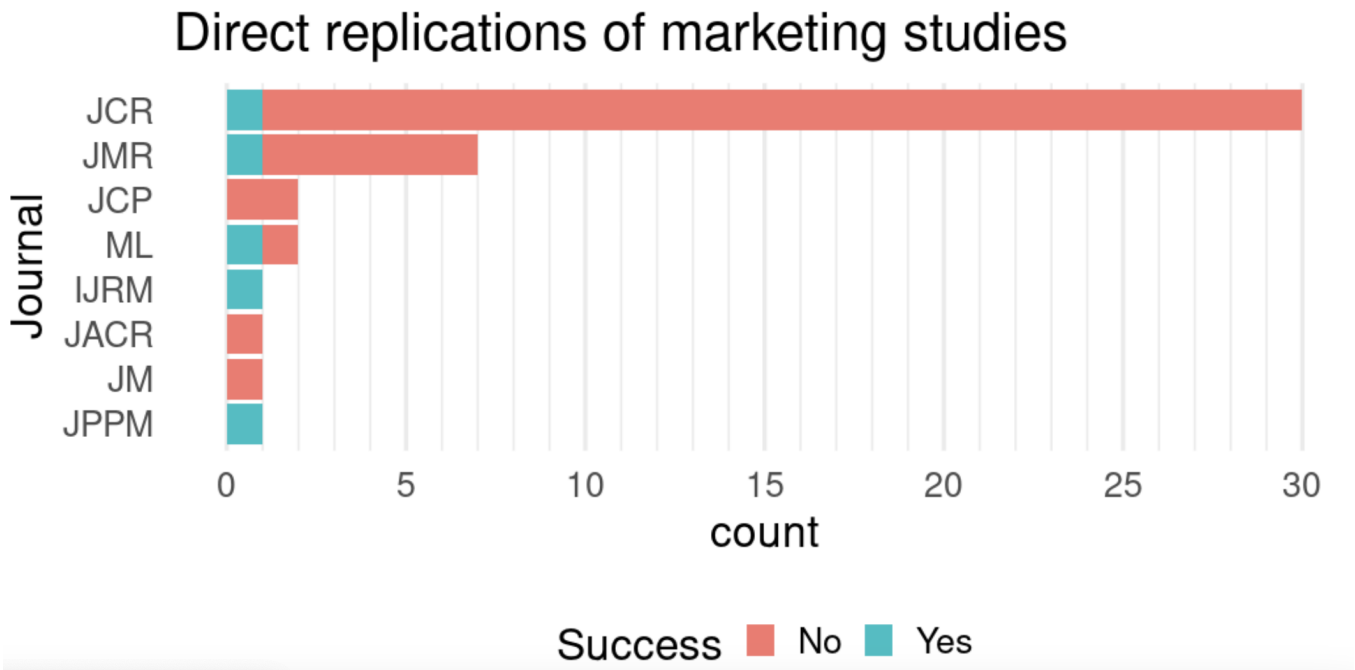
- This pattern suggests a strong **publication bias** in favor of significance, possibly driven by incentives to publish.

Replication Crisis

- **Cornerstone of Science:** Replication of experiments.

Replication Crisis

- **Cornerstone of Science:** Replication of experiments.
  - **Why?** Confirms reliability, detects false positives and P-hacking, transparency and accountability, etc.



- Many replications in marketing studies **fail**, highlighting concerns about reproducibility in published research.

P-hacking

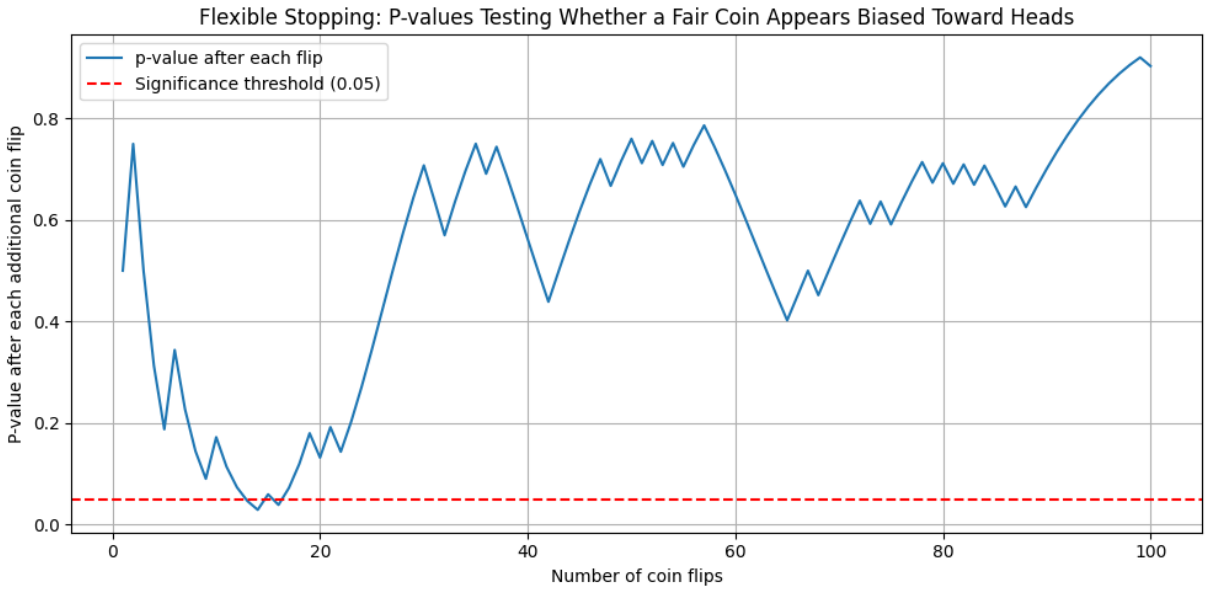
- **P-hacking:** Producing false positives by being flexible in experimental design and data analysis.
- This flexibility often leads to misleading results and contributes significantly to the replication crisis.

Forms of P-hacking

1. Flexible Stopping
2. HARKing (Hypothesizing after reulsts are known)
3. Outlier Removal

Forms of P-hacking – (1) Flexible Stopping

- **Flexible Stopping:**
  - Adding data one observation at a time and repeatedly testing for significance until the p-value meets the threshold.
  - Each additional test increases the likelihood of obtaining a false positive.
  - Performing many tests (e.g., comparing subgroups) inflates the chances of finding at least one significant result.
  - Reporting only the significant results, while ignoring the non-significant ones, makes the findings appear valid.



Forms of P-hacking – (2) HARKing (Hypothesizing After Results Are Known)

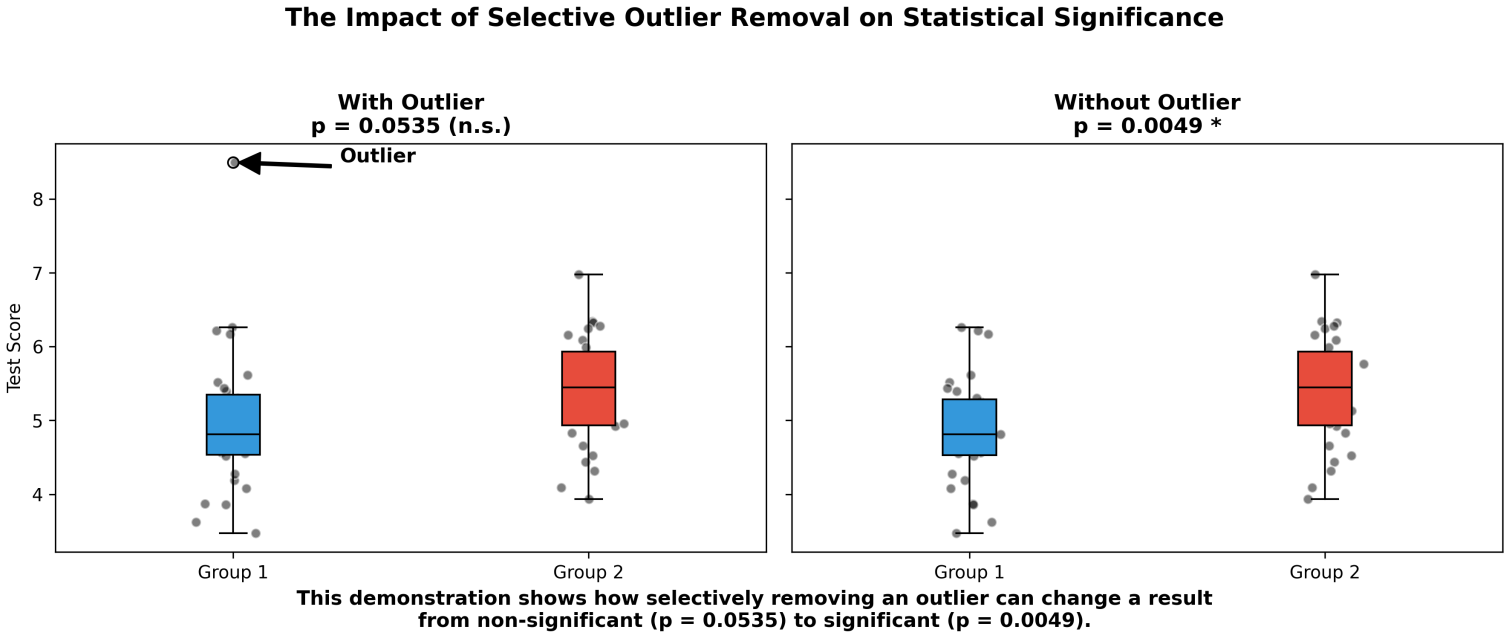
- Testing multiple dependent variables and developing hypotheses post-hoc based on significant findings.
- Reporting only the significant results as if they were the original hypotheses.
- **Analogy:** Throwing a dart and then painting a bullseye around where it lands.

Argument:

- The p-value assumes a single, pre-specified hypothesis. Post-hoc results inflate the false positive rate and often fail to replicate.

Forms of P-hacking – (3) Outlier Removal

- **When It's Okay:** Removing clearly erroneous data points before any analysis begins (e.g., sensor malfunctions, data entry errors).
- **When It's Not Okay:** Looking at the results first and selectively removing data points to make results significant. This practice is often referred to as "peeking".



Solutions to P-hacking

- Addressing the causes of the replication crisis requires a combination of methodological and cultural changes in research.
- Some approaches include:

1. **Pre-registration:** Researchers specify their hypotheses, data collection plans, and analysis methods before conducting the study.
2. **Lowering α to 0.005:** Reduces the likelihood of false positives but may require larger sample sizes.
3. **Report Effect Size and Confidence Intervals**
4. **Increase Power**

Effect Size (Cohen's d)

1. **Statistical Significance (p-value):** Tells us whether an effect is likely due to chance, but **does not indicate** the size or importance of the effect.
   - p-value $\rightarrow$ Statistical Significance
2. **Effect Size (Cohen's d):** Quantifies how large or meaningful the effect is – i.e., "How much does it matter?"
   - Cohen's d $\rightarrow$ Practical Significance

Cohen's d (Effect Size Formula)

$$\text{Cohen's } d = \frac{\mu_1 - \mu_2}{\sigma}$$

- $\mu_1 - \mu_2$ is the difference between group means.
- σ is the pooled or within-group standard deviation.

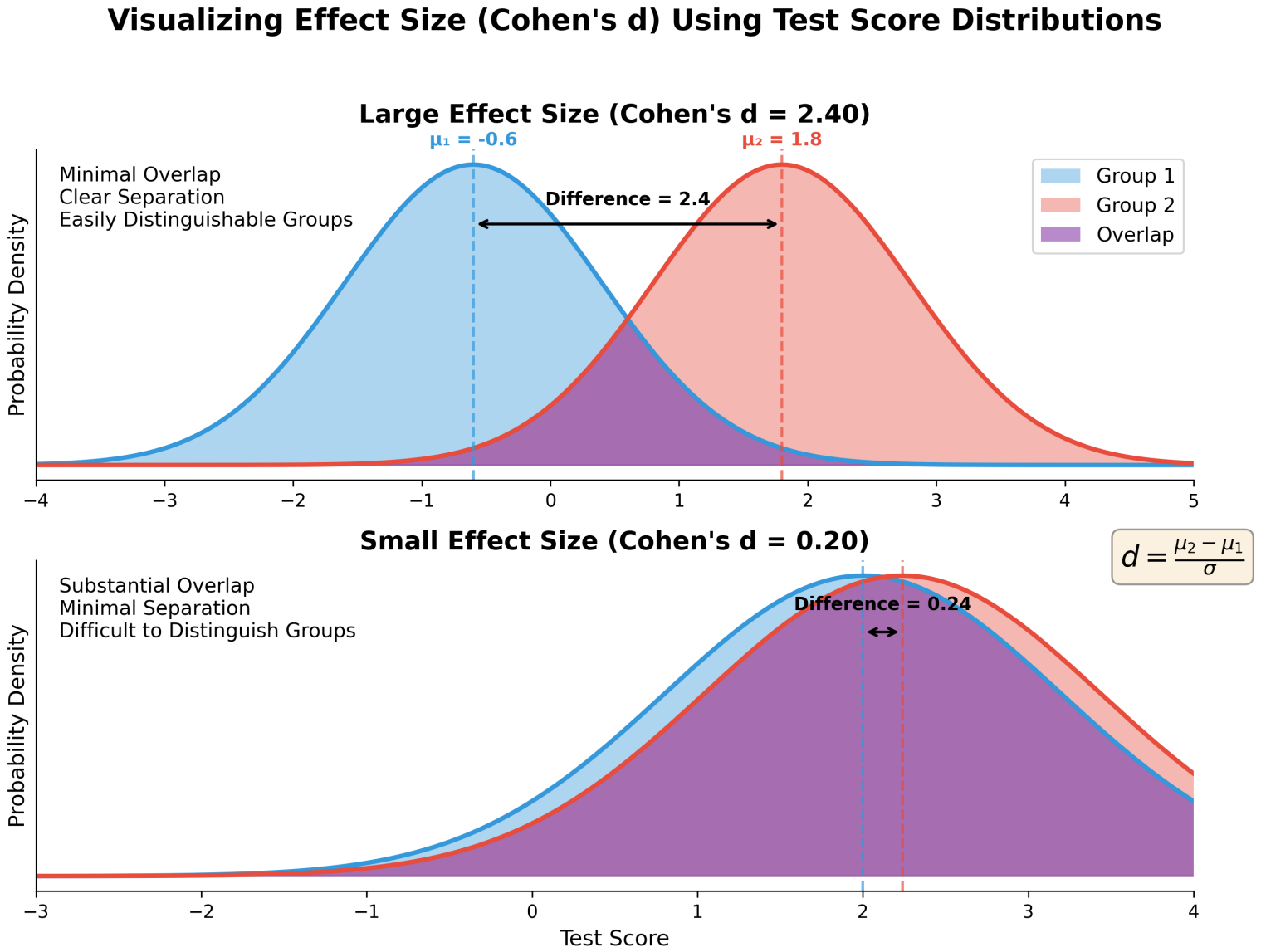
Cohen's d Contextulaization

Size	Effect size	Example (from Cohen 1969)
'Large'	0.8	difference between heights of 13- and 18-year-old girls in the US
'Mediu m'	0.5	difference between heights of 14- and 18-year-old girls in the US
'Small'	0.2	difference between heights of 15- and 16-year-old girls in the US

Range:

- **Technically**, $(-\infty, \infty)$.
- **Practically**, most values are between 0-1.

Effect Size Visualization



What Affects Effect Size?

Cohen's d (Effect Size Formula)

$$\text{Cohen's } d = \frac{\mu_1 - \mu_2}{\sigma}$$

1. **Smaller Mean Differences:** Smaller mean differences lead to smaller effect sizes.
2. **Larger Standard Deviation:** Greater variability (standard deviation) reduces the effect size.
3. **Sample Size and Effect Size:** While increasing the sample size boosts statistical power (the ability to detect an effect), it does not change the effect size itself.

Why Does Effect Size Matter? – Example

- **Suppose** a new drug shows a statistically significant effect (i.e., a small p-value) in helping cure a disease.
- **However,** the effect size is very small –the improvement is minimal in practical terms.
- As a hospital, it may not be worth switching manufacturers, overhauling the supply chain, or risking new side effects for such a minor gain.

Power

- **Question:** How can we increase the probability of observing an effect size, if it exists?

Power

- **Question:** How can we increase the probability of observing an effect size, if it exists?
- Increase statistical **power**!

		Reality	
		Effect	No Effect
Significant?	Yes	Correct (1-β)	Type I Error False Positive (α)
	No	Type II Error False Negative (β)	Correct (1-α)

Table: Confusion Matrix

- **Power:** The probability of detecting a real effect if it exists.
- **Formally,** the probability of **not** making a Type II error

$$\text{Power} = P(1 - \beta)$$

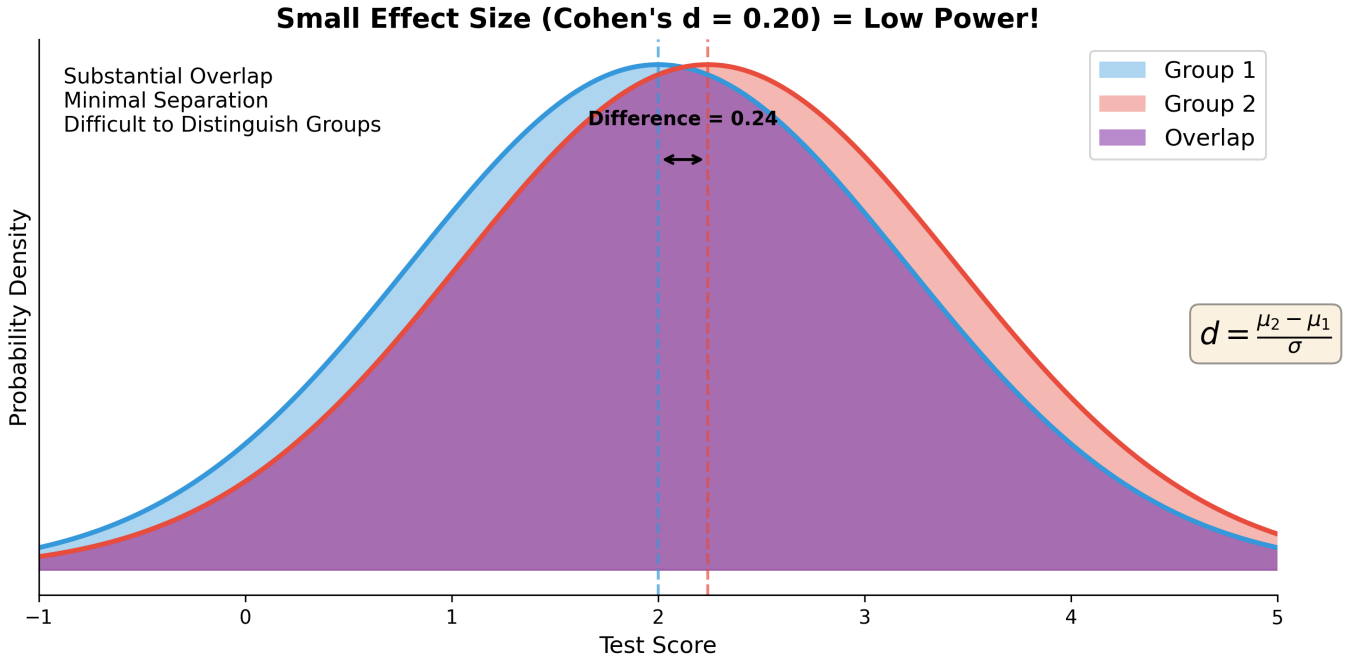
What Affects Power?

- The statistical power of a test depends on several key factors:

1. **Sample Size:** Larger sample size $\Rightarrow$ higher power.
2. **Effect Size:** Larger effect size (e.g., greater group mean difference) $\Rightarrow$ higher power.
3. **Significance Level (α):** Lowering α (e.g., from 0.05 to 0.005) makes the test more conservative $\Rightarrow$ lower power.
4. **Type of Test:** Parametric tests have higher power than non-parametric ones, assuming assumptions are met.
5. **Standard Deviation (σ):** Smaller pooled variability $\Rightarrow$ higher power.

Relationship Between Effect Size and Power

- Effect size and statistical power are intrinsically linked.
- **Effect size** quantifies the magnitude of a relationship or difference in the data
- **Statistical power** reflects the ability to detect that effect.



- We need a lot of data to detect such a small effect. That's the power of big data!

Power Calculation and The Replicability Crisis as a Power Failure

- Since power depends on many factors, there is no formula for us to calculate power. We estimate it from simulation.
- G*Power (clickable link!): A free software that does the simulation for you.

Power Calculation and The Replicability Crisis as a Power Failure

- Since power depends on many factors, there is no formula for us to calculate power. We estimate it from simulation.
- G*Power (clickable link!): A free software that does the simulation for you.

- The replicability crisis in research can be seen as a direct consequence of low statistical power:

1. Starting with a small ***sample size*** $\rightarrow$ low power $\rightarrow$ "negative" result $\rightarrow$ hack it to publish $\rightarrow$ false positive $\rightarrow$ failure to replicate.
2. Starting with a small ***effect size*** $\rightarrow$ low power $\rightarrow$ "negative" result $\rightarrow$ hack it to publish $\rightarrow$ false positive $\rightarrow$ failure to replicate.

Summary: P-value, Effect Size, and Power

Metric	Interpretation
P-value	Statistical Significance: Is the effect unlikely due to chance?
Effect Size	Practical Significance: How large or meaningful is the effect?
Power	Ability to detect an effect if one truly exists