From one environment to many: The problem of replicability of statistical inferences

Michael J. Higgins

Work with James J. Higgins and Jinguang Lin

Kansas State University

Aug. 1, 2019

Reproducibility crisis: Examples

- Recent problem that many scientific findings cannot be replicated.

  Replicability: "The ability of a researcher to duplicate the results of a prior study if the same procedures are followed but new data are collected."

Reproducibility crisis: Examples

- Recent problem that many scientific findings cannot be replicated.

  Replicability: "The ability of a researcher to duplicate the results of a prior study if the same procedures are followed but new data are collected."
- Problem exists in nearly all fields of science: **Reproducibility crisis**.

Reproducibility crisis: Examples

- Recent problem that many scientific findings cannot be replicated.

  Replicability: "The ability of a researcher to duplicate the results of a prior study if the same procedures are followed but new data are collected."
- Problem exists in nearly all fields of science: **Reproducibility crisis**.
- A 2016 survey of *Nature* readers found that **about 70%** of scientists have failed to reproduce other researchers' experiments, and **more than 50%** have failed to reproduce their own studies.

Reproducibility crisis: Examples

- Recent problem that many scientific findings cannot be replicated.

  Replicability: "The ability of a researcher to duplicate the results of a prior study if the same procedures are followed but new data are collected."
- Problem exists in nearly all fields of science: **Reproducibility crisis**.
- A 2016 survey of *Nature* readers found that **about 70%** of scientists have failed to reproduce other researchers' experiments, and **more than 50%** have failed to reproduce their own studies.
- In 2015, a study in Science had researchers conduct replications of 100 experimental and correlational studies published in three psychology journals using original materials.

  Less than 50% of the original findings were replicable.

Reproducibility crisis: Response

- Concerns about replicability has lead to a large backlash against terms like p-values and statistical significance.

Reproducibility crisis: Response

- Concerns about replicability has lead to a large backlash against terms like *p*-values and statistical significance.
- In 2015, the journal *Basic and Applied Social Psychology* **banned the use of *p*-values** and claims of statistical significance.

Reproducibility crisis: Response

- Concerns about replicability has lead to a large backlash against terms like *p*-values and statistical significance.
- In 2015, the journal *Basic and Applied Social Psychology* **banned the use of *p*-values** and claims of statistical significance.
- A 2019 special issue of *The American Statistican* recommended that the use of the phrase **"statistically significant" be abandoned completely**.

Reproducibility crisis: Causes

- What are the causes of the reproducibility crisis?

Reproducibility crisis: Causes

- What are the causes of the reproducibility crisis?
- *p*-hacking: manipulating and dredging through data until a significant result is found.

  E.g.: No significant average treatment effect, but results suggest that treatment is effective on left-handed, blonde, 38 year olds.

Reproducibility crisis: Causes

- What are the causes of the reproducibility crisis?
- *p*-hacking: manipulating and dredging through data until a significant result is found.

  E.g.: No significant average treatment effect, but results suggest that treatment is effective on left-handed, blonde, 38 year olds.
- Overstatement of power.

  E.g.: Performing a cluster randomized trial, but analyzing experiments assuming treatment was given to observational units.

Reproducibility crisis: Causes

- What are the causes of the reproducibility crisis?
- *p*-hacking: manipulating and dredging through data until a significant result is found.

  E.g.: No significant average treatment effect, but results suggest that treatment is effective on left-handed, blonde, 38 year olds.
- Overstatement of power.

  E.g.: Performing a cluster randomized trial, but analyzing experiments assuming treatment was given to observational units.
- Ignoring multiple testing issues.

Reproducibility crisis: Causes

- What are the causes of the reproducibility crisis?
- *p*-hacking: manipulating and dredging through data until a significant result is found.

  E.g.: No significant average treatment effect, but results suggest that treatment is effective on left-handed, blonde, 38 year olds.
- Overstatement of power.

  E.g.: Performing a cluster randomized trial, but analyzing experiments assuming treatment was given to observational units.
- Ignoring multiple testing issues.
- Inaccurate protocol/statistical code described to obtain result.

  E.g.: Missing details on how missing data were imputed, whether observations were eliminated, if/how data were transformed, etc.

Environment by treatment interaction

- Most of the aforementioned issues pointed at researcher ignorance, honest mistakes, or deliberate fraud as primary culprits of the reproducibility crisis: **researcher's fault**.

Environment by treatment interaction

- Most of the aforementioned issues pointed at researcher ignorance, honest mistakes, or deliberate fraud as primary culprits of the reproducibility crisis: **researcher's fault**.
- **Our claim**: A study may fail to be replicable **even if both the initial and follow-up studies follow best statistical practices**.

Environment by treatment interaction

- Most of the aforementioned issues pointed at researcher ignorance, honest mistakes, or deliberate fraud as primary culprits of the reproducibility crisis: **researcher's fault**.
- **Our claim**: A study may fail to be replicable **even if both the initial and follow-up studies follow best statistical practices**.
- **The culprit**: Environment by treatment interaction.

Environment by treatment interaction

- Most of the aforementioned issues pointed at researcher ignorance, honest mistakes, or deliberate fraud as primary culprits of the reproducibility crisis: **researcher's fault**.
- **Our claim**: A study may fail to be replicable **even if both the initial and follow-up studies follow best statistical practices**.
- **The culprit**: Environment by treatment interaction.
- Statistical inferences that are drawn from the analysis of data apply only to the environment in which the study is conducted. Often, researchers want their result to apply more broadly.

Environment by treatment interaction

- Most of the aforementioned issues pointed at researcher ignorance, honest mistakes, or deliberate fraud as primary culprits of the reproducibility crisis: **researcher's fault**.
- **Our claim**: A study may fail to be replicable **even if both the initial and follow-up studies follow best statistical practices**.
- **The culprit**: Environment by treatment interaction.
- Statistical inferences that are drawn from the analysis of data apply only to the environment in which the study is conducted. Often, researchers want their result to apply more broadly.
- **We show**: The presence of differing treatment effects across environments can dramatically decrease the likelihood of replicating results.

Environment by treatment interaction

- Most of the aforementioned issues pointed at researcher ignorance, honest mistakes, or deliberate fraud as primary culprits of the reproducibility crisis: **researcher's fault**.
- **Our claim**: A study may fail to be replicable **even if both the initial and follow-up studies follow best statistical practices**.
- **The culprit**: Environment by treatment interaction.
- Statistical inferences that are drawn from the analysis of data apply only to the environment in which the study is conducted. Often, researchers want their result to apply more broadly.
- **We show**: The presence of differing treatment effects across environments can dramatically decrease the likelihood of replicating results.
- Worst case: replicating a study is **equivalent to flipping a fair coin**.

Goal of talk

- How can we assess how environment by treatment interaction can impact replicability of a result?

Goal of talk

- How can we assess how environment by treatment interaction can impact replicability of a result?
- If researchers are aware of this interaction, how can inferences be adjusted?

Goal of talk

- How can we assess how environment by treatment interaction can impact replicability of a result?
- If researchers are aware of this interaction, how can inferences be adjusted?
- **Measures of replicability**:
  1. Probability of replicability
  2. Adjusted p-value
  3. Adjusted confidence interval

Model

Observations from the initial study follow **Model 1**:

$$Y_{ij} = \mu_i + \epsilon_{ij}, \quad i = 1, 2, \quad j = 1, ..., n \tag{1}$$

Observations from the follow-up study follow **Model 2**:

$$Y_{ij} = \mu_i + \theta + \delta_i + \epsilon_{ij}, \quad i = 1, 2, \quad j = 1, ..., n \tag{2}$$

- μ_i: mean for treatment i. Assume $\mu_1 > \mu_2$.
- ϵ_{ij}: i.i.d. $N(0, \sigma_e^2)$: Random error terms.
- θ: $N(0, \sigma_\theta^2)$: Random environment effect affecting both treatments.
- δ_i: i.i.d. $N(0, \sigma_B^2)$: Random environment effect affecting only treatment i.

Model

Observations from the initial study follow **Model 1**:

$$Y_{ij} = \mu_i + \epsilon_{ij}, \quad i = 1, 2, \quad j = 1, ..., n \tag{1}$$

Observations from the follow-up study follow **Model 2**:

$$Y_{ij} = \mu_i + \theta + \delta_i + \epsilon_{ij}, \quad i = 1, 2, \quad j = 1, ..., n \tag{2}$$

- μ_i: mean for treatment i. Assume $\mu_1 > \mu_2$.
- ϵ_{ij}: i.i.d. $N(0, \sigma_e^2)$: Random error terms.
- θ: $N(0, \sigma_\theta^2)$: Random environment effect affecting both treatments.
- δ_i: i.i.d. $N(0, \sigma_B^2)$: Random environment effect affecting only treatment i.
- Environment: Includes factors such as location/facility, time, personnel, equipment. Certain aspects may be uncontrollable.

Model

Observations from the follow-up study follow **Model 2**:

$$Y_{ij} = \mu_i + \theta + \delta_i + \epsilon_{ij}, \quad i = 1, 2, \quad j = 1, ..., n$$

Under Model 2, the difference of sample means can be expressed as

$$\overline{Y}_1 - \overline{Y}_2 = (\mu_1 - \mu_2) + (\overline{\epsilon}_1 - \overline{\epsilon}_2) + (\delta_1 - \delta_2)$$

Thus, $\overline{Y}_1 - \overline{Y}_2$ has a normal distribution with mean $\mu_1 - \mu_2$ and variance $2\sigma_e^2/n + 2\sigma_B^2$.

Two-sample t-test

We assume that inferences are performed on $\mu_1 - \mu_2$ using a two-sample t-test.

$$T = \frac{\overline{Y}_1 - \overline{Y}_2}{S_p\sqrt{2/n}} = \frac{(\overline{Y}_1 - \overline{Y}_2)}{S_p\sqrt{2/n}\sqrt{2\sigma_e^2/n + 2\sigma_B^2}}\sqrt{2\sigma_e^2/n + 2\sigma_B^2}$$

$$= NCT\sqrt{1 + n\sigma_B^2/\sigma_e^2}$$

- S_p: the "pooled" sample standard deviation
- NCT: non-central t-distribution with $df = 2(n-1)$ and non-centrality parameter

$$u = \frac{\mu_1 - \mu_2}{\sqrt{2\sigma_e^2/n + 2\sigma_B^2}} = \frac{\sqrt{n}\Delta}{\sqrt{1 + n\sigma_B^2/\sigma_e^2}}$$

Effect Size, EER

NCT: non-central t-distribution with $df = 2(n-1)$ and non-centrality parameter

$$u = \frac{\mu_1 - \mu_2}{\sqrt{2\sigma_e^2/n + 2\sigma_B^2}} = \frac{\sqrt{n}\Delta}{\sqrt{1 + n\sigma_B^2/\sigma_e^2}}$$

- Δ is the **effect size**:

$$\Delta = (\mu_1 - \mu_2)/(\sigma_e\sqrt{2}).$$

- σ_B/σ_e is the **environmental effect ratio (EER)**: The ratio of the standard deviation between environments within treatment (or interaction) and the standard deviation of experimental error.

Effect Size, EER

NCT: non-central t-distribution with $df = 2(n-1)$ and non-centrality parameter

$$u = \frac{\mu_1 - \mu_2}{\sqrt{2\sigma_e^2/n + 2\sigma_B^2}} = \frac{\sqrt{n}\Delta}{\sqrt{1 + n\sigma_B^2/\sigma_e^2}}$$

- Δ is the **effect size**:

$$\Delta = (\mu_1 - \mu_2)/(\sigma_e\sqrt{2}).$$

- σ_B/σ_e is the **environmental effect ratio (EER)**: The ratio of the standard deviation between environments within treatment (or interaction) and the standard deviation of experimental error.
- Δ and the EER are **two critical factors** in assessing the replicability of a study.

Probability of replicability

- **Probability of replicability**: The probability that the follow-up study yields a significant result, assuming the initial study was significant.

Probability of replicability

- **Probability of replicability**: The probability that the follow-up study yields a significant result, assuming the initial study was significant.
- If $T \geq t_{\alpha/2,df}$ in the initial study, then we have a replicable result if $T \geq t_{\alpha/2,df}$ in the follow-up experiment.

  ($t_{\alpha,df}$ is the $1-\alpha$ quantile of the t-distribution).

Probability of replicability

- **Probability of replicability**: The probability that the follow-up study yields a significant result, assuming the initial study was significant.
- If $T \geq t_{\alpha/2,df}$ in the initial study, then we have a replicable result if $T \geq t_{\alpha/2,df}$ in the follow-up experiment.

  ($t_{\alpha,df}$ is the $1-\alpha$ quantile of the t-distribution).
- If the initial study is significant in the wrong direction—that is, $T \leq -t_{\alpha/2,df}$ in the initial study—then replicability occurs if $T \leq -t_{\alpha/2,df}$ in the follow-up study.

  This is confirmation of an incorrect result.

Probability of replicability

- For a two-sided test at level of significance α, the probability of replicability is approxmiately

$$1 - G_{df,u}\left(t_{\alpha/2,df}/\sqrt{1 + n\sigma_B^2/\sigma_e^2}\right)$$

- $G_{df,u}(t)$: the c.d.f. of the non-central t-distribution, with non-centrality parameter u.

Probability of replicability

- For a two-sided test at level of significance α, the probability of replicability is approxmiately

$$1 - G_{df,u}\left(t_{\alpha/2,df}/\sqrt{1 + n\sigma_B^2/\sigma_e^2}\right)$$

- $G_{df,u}(t)$: the c.d.f. of the non-central t-distribution, with non-centrality parameter u.
- This probability depends on:
  1. n, α, as usual.
  2. Δ: Larger $\Delta \rightarrow$ Larger probability of replicability.
  3. EER: Smaller EER $\rightarrow$ Larger probability of replicability (if original study has high power).

Example: Probability of replicability

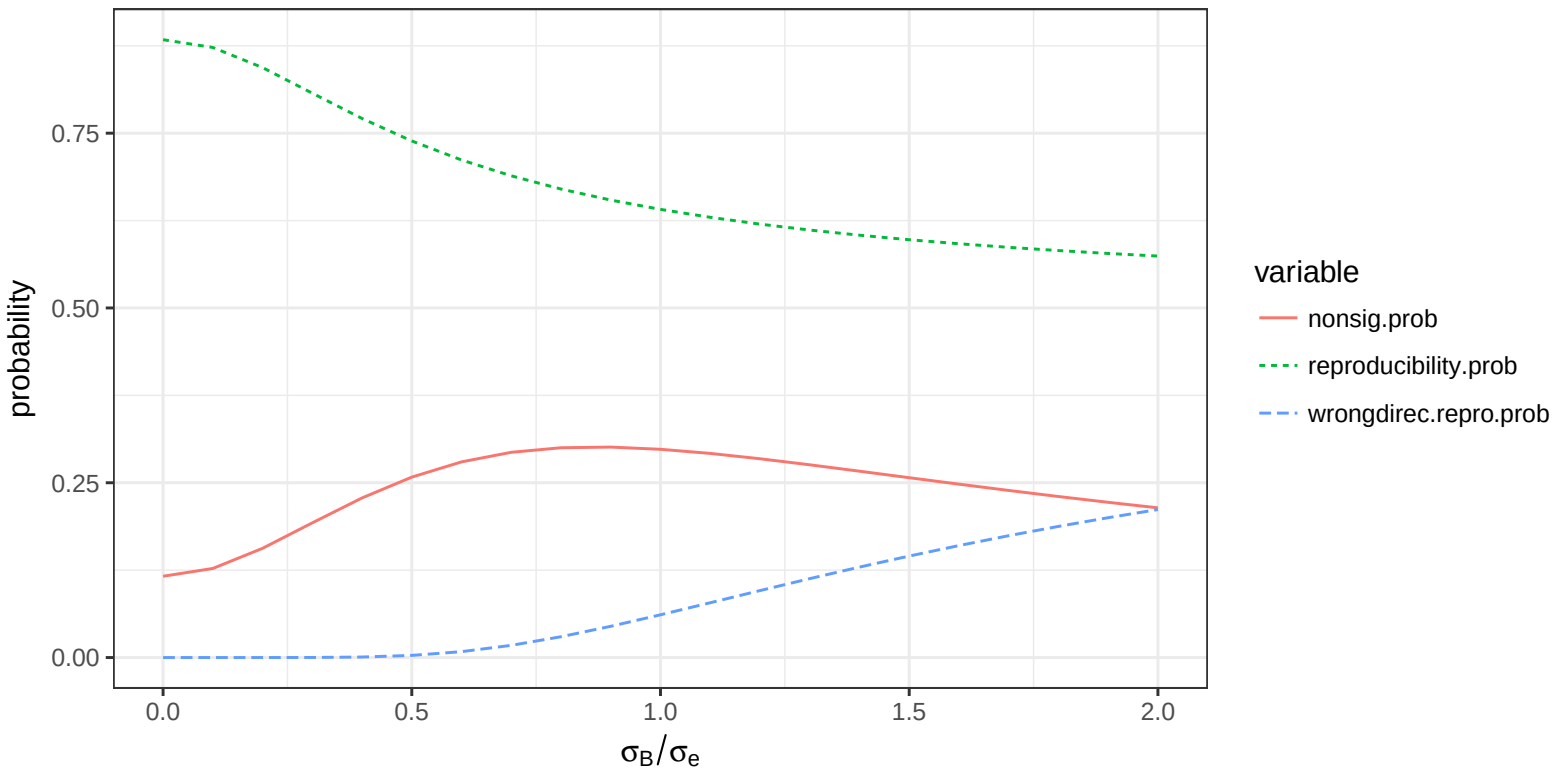

Figure: Probability of replicability, probability of significance in the wrong direction, non-significance, vs. σ_B/σ_e, $n = 11$, Δ=1.02, $\alpha = .05$, initial power = .88

Adjusted p-value

- Consider a hypothesis test:

$$H_0 : \mu_1 - \mu_2 = 0 \quad \text{against} \quad H_1 : \mu_1 - \mu_2 \neq 0$$

- The observed effect size: $\Delta^* = (\overline{y}_1 - \overline{y}_2)/(\sqrt{2}S_p)$.
- The observed t-statistic: $T = \dfrac{\overline{y}_1 - \overline{y}_2}{S_p\sqrt{2/n}} = \Delta^*\sqrt{n}$

Adjusted p-value

- Consider a hypothesis test:

$$H_0 : \mu_1 - \mu_2 = 0 \quad \text{against} \quad H_1 : \mu_1 - \mu_2 \neq 0$$

- The observed effect size: $\Delta^* = (\overline{y}_1 - \overline{y}_2)/(\sqrt{2}S_p)$.
- The observed t-statistic: $T = \dfrac{\overline{y}_1 - \overline{y}_2}{S_p\sqrt{2/n}} = \Delta^*\sqrt{n}$
- Under the null hypothesis, the non-centrality parameter $u = 0$. The two-sided p-value for the observed t-stat under Model 2 is:

$$2\left(1 - G_{df,u=0}\left(\Delta^*\sqrt{n}/\sqrt{1 + n\sigma_B^2/\sigma_e^2}\right)\right)$$

- This is called the **adjusted p-value**.

Adjusted p-value

Properties of the adjusted p-value:

- The adjusted p-value is smaller for larger effect size Δ and smaller EER.

Adjusted p-value

Properties of the adjusted p-value:

- The adjusted p-value is smaller for larger effect size Δ and smaller EER.
- Can invert the test using the adjusted p-value to form an adjusted confidence interval.

Adjusted p-value

Properties of the adjusted p-value:

- The adjusted p-value is smaller for larger effect size Δ and smaller EER.
- Can invert the test using the adjusted p-value to form an adjusted confidence interval.
- For given values of Δ and EER, **it may be impossible** to achieve an adjusted p-value < 0.05, **regardless of the sample size**.

Adjusted p-value

Properties of the adjusted p-value:

- The adjusted p-value is smaller for larger effect size Δ and smaller EER.
- Can invert the test using the adjusted p-value to form an adjusted confidence interval.
- For given values of Δ and EER, **it may be impossible** to achieve an adjusted p-value < 0.05, **regardless of the sample size**.
- Previous solutions for increasing replicability, such as drawing larger sample sizes or using smaller significance levels, do not account for variability across environments.

Adjusted p-value

Properties of the adjusted p-value:

- The adjusted p-value is smaller for larger effect size Δ and smaller EER.
- Can invert the test using the adjusted p-value to form an adjusted confidence interval.
- For given values of Δ and EER, **it may be impossible** to achieve an adjusted p-value < 0.05, **regardless of the sample size**.
- Previous solutions for increasing replicability, such as drawing larger sample sizes or using smaller significance levels, do not account for variability across environments.
- Large effect size is a **better indicator of replicability** than a large sample size n or a small traditional p-value.

Adjusted p-value: Example

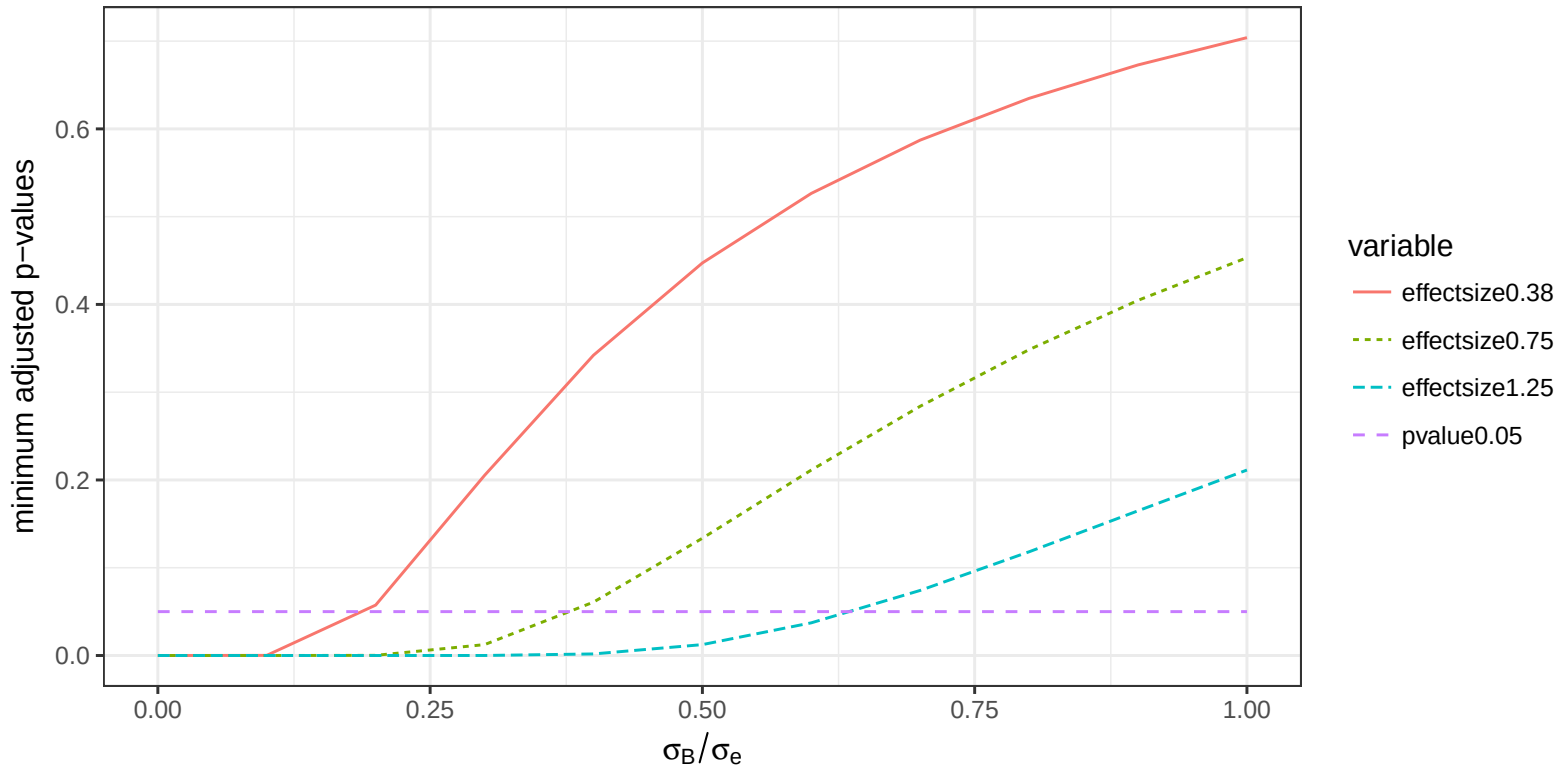

Figure: Adjusted p-values vs. EER for various effect sizes Δ, $n \to \infty$.

How to determine EER

- In our experience: Common values of EER range from 0.2 to 1.0, but can be lower or higher depending on the setting. Larger effect sizes are needed to overcome larger values of EER.

How to determine EER

- In our experience: Common values of EER range from 0.2 to 1.0, but can be lower or higher depending on the setting. Larger effect sizes are needed to overcome larger values of EER.
- Empirical estimate of EER: Perform a study at separate labs simultaneously. Can get an estimate of EER directly (though can be quite costly).

How to determine EER

- In our experience: Common values of EER range from 0.2 to 1.0, but can be lower or higher depending on the setting. Larger effect sizes are needed to overcome larger values of EER.
- Empirical estimate of EER: Perform a study at separate labs simultaneously. Can get an estimate of EER directly (though can be quite costly).
- Historical estimate: A meta-analysis may allow a good estimate of EER for some types of studies.

How to determine EER

- In our experience: Common values of EER range from 0.2 to 1.0, but can be lower or higher depending on the setting. Larger effect sizes are needed to overcome larger values of EER.
- Empirical estimate of EER: Perform a study at separate labs simultaneously. Can get an estimate of EER directly (though can be quite costly).
- Historical estimate: A meta-analysis may allow a good estimate of EER for some types of studies.
- Sensitivity analysis: Statistical analysis may include values of EER for which the given result will replicate.

Thank you

Thank you!

Paper available at:

https://arxiv.org/abs/1904.10036