Assessing Replication in Education Research

Jake Schauer | Department of Statistics | Northwestern University
jms@u.northwestern.edu

THE REPLICATION CRISIS

Replicating scientific findings raises important questions about the body of knowledge in education. Research that spans different schools, districts, or states can help identify findings that are robust to setting or implementation. It also provides greater insight on what works for whom and why. At its core, replications–whether conceptual or precise–shore up the evidence on which we base important policy decisions. Especially at a time of shaken confidence in the replicability of findings in other fields–medicine and psychology, for example–it is important to understand what findings in education stand up to this scrutiny.

PARAMETERS OF REPLICATION

For each replicate, we have a true effect size θ_i that is estimated by T_i with sampling variance v_i. Inferences about replication are inferences about how similar the θ_i are. Here, we characterize their differences by τ^2, the variance of a distribution from which the θ_i were drawn.

Study Results

Given study i, we observe T_i the estimate of the treatment effect, and v_i its sampling variance. We assume that each T_i is normally distributed around the true treatment effect θ_i with variance v_i. **These results T_i can tell us about the θ_i and τ^2.**

Effect Parameters

Without knowing how the true effect parameters θ_i may differ, we can model them as draws from a distribution. This is equivalent to a random effects model in meta-analysis (Hedges & Vevea, 1998). If the distribution would generate θ_i that are very similar, we would conclude that the finding replicates; **inferences about replication are inferences about the distribution from which the θ_i are drawn.**

Heterogeneity

Differences among the study results can be characterized by τ^2, the variance of the distribution that generated them. Our conclusions about replication will depend on the magnitude of τ^2. How large of a value of τ^2 corresponds to approximate replication is a matter of scientific judgment. Here we consider τ^2 relative to the observed sampling variances v_i, and use the conventions from different scientific fields (Hedges & Pigott, 2001):

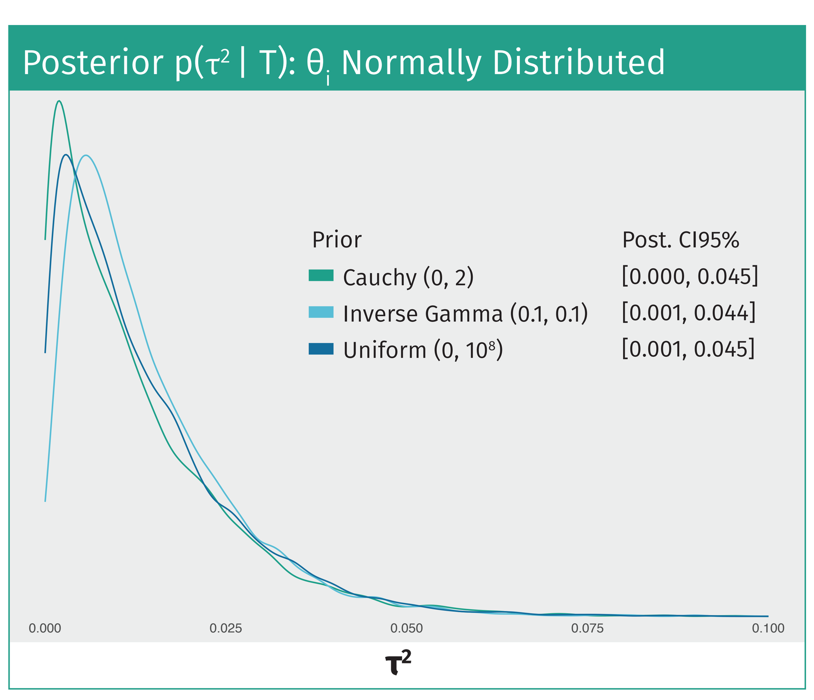
- **$\tau^2 = 0$: exact replication**
- **$\tau^2 \le v/4$: approximate replication (physics)**
- **$\tau^2 \le v/3$: approximate replication (personnel psychology)**
- **$\tau^2 \le 2v/3$: approximate replication (medicine)**

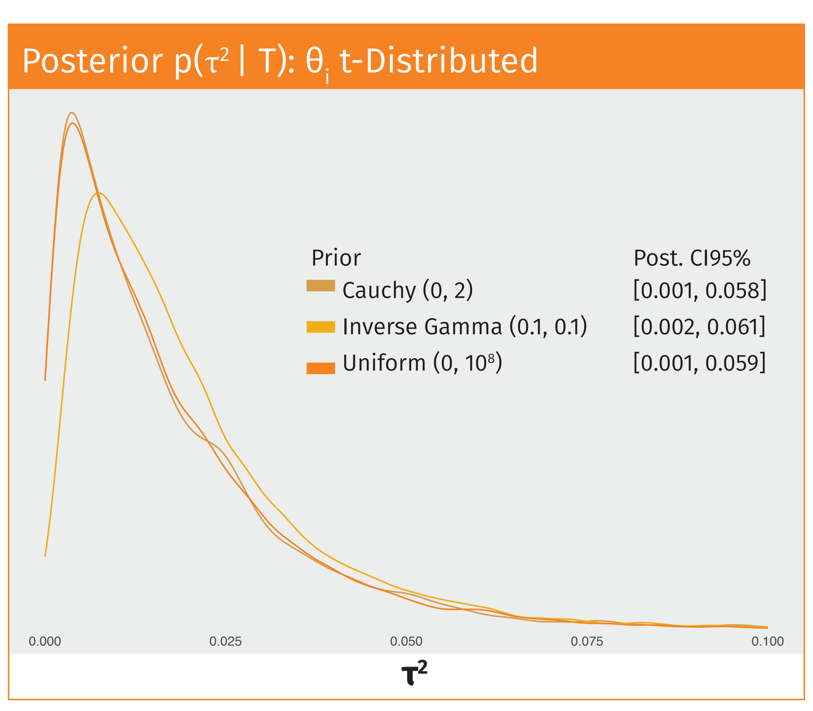
There is no established convention in education!

PRIOR SPECIFICATION

We use Cauchy, inverse-gamma, and uniform prior distributions, which all reflect the fact that we know nothing about replication *a priori*.

These prior distributions all produce similar poster distributions in that they have a similar shape and nearly identical 95% posterior credible intervals for τ^2.



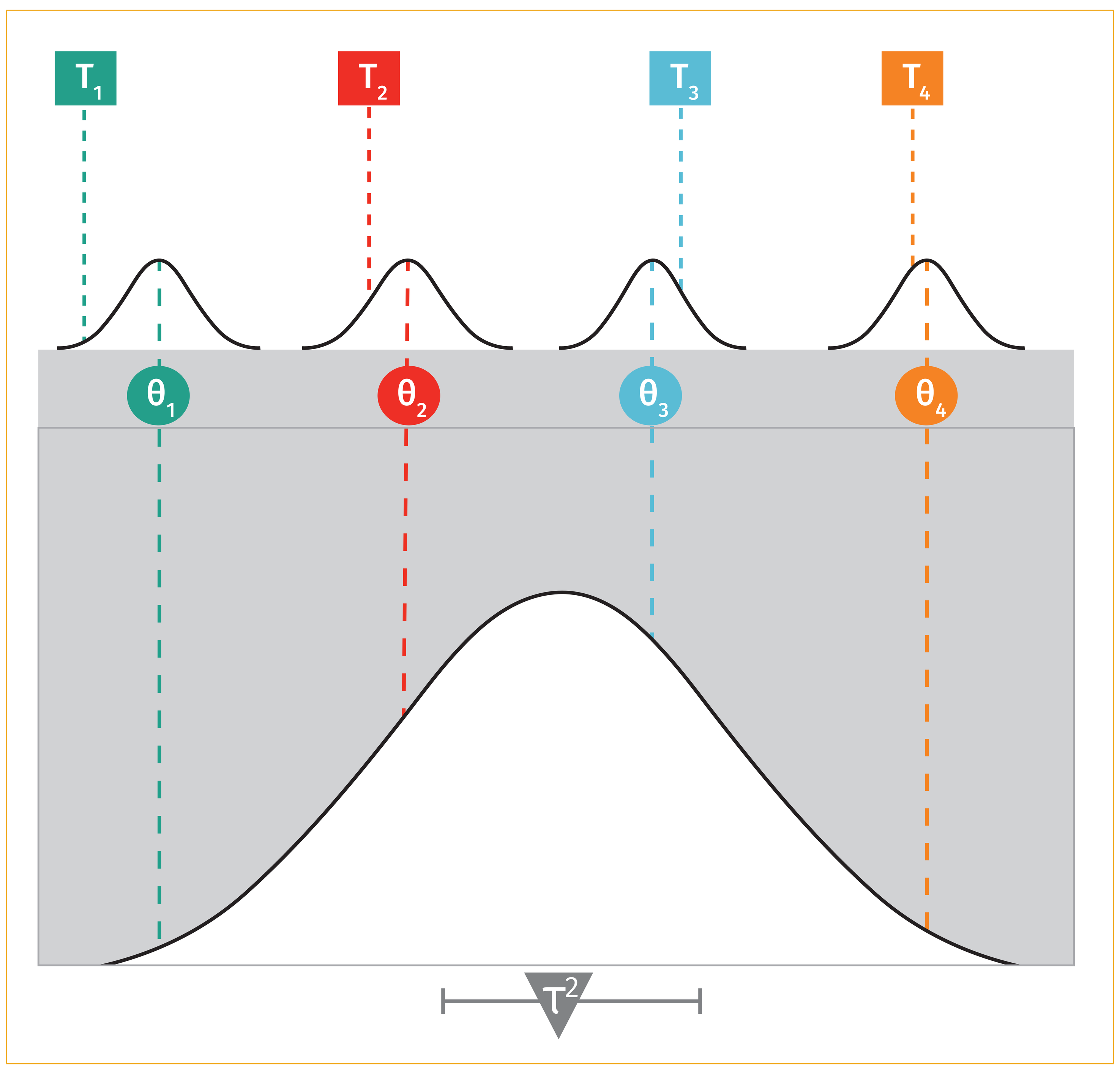


WHY DON'T STUDIES REPLICATE?

Studies may get different results due to:

Statistical issues	SAMPLING ERROR	Arises from conducting experiments on a sample rather than the entire population.
Replication issues	POPULATION DIFFERENCES	Nonignorable sample selection: the populations for each study are different. E.g., one study takes place in Arizona and the other in Minnesota.
	TREATMENT DIFFERENCES	Due to conditions on the ground, interventions may need to be altered from study to study. E.g., in-class interventions may vary in frequency due to resources.
	EXPERIMENTAL ERROR	Experimental design, poor measurement, attrition, and other issues–if they are correlated with the treatment effect–may lead to different conclusions.

DATA GENERATING PROCESS



EXAMPLE: MANYLABS

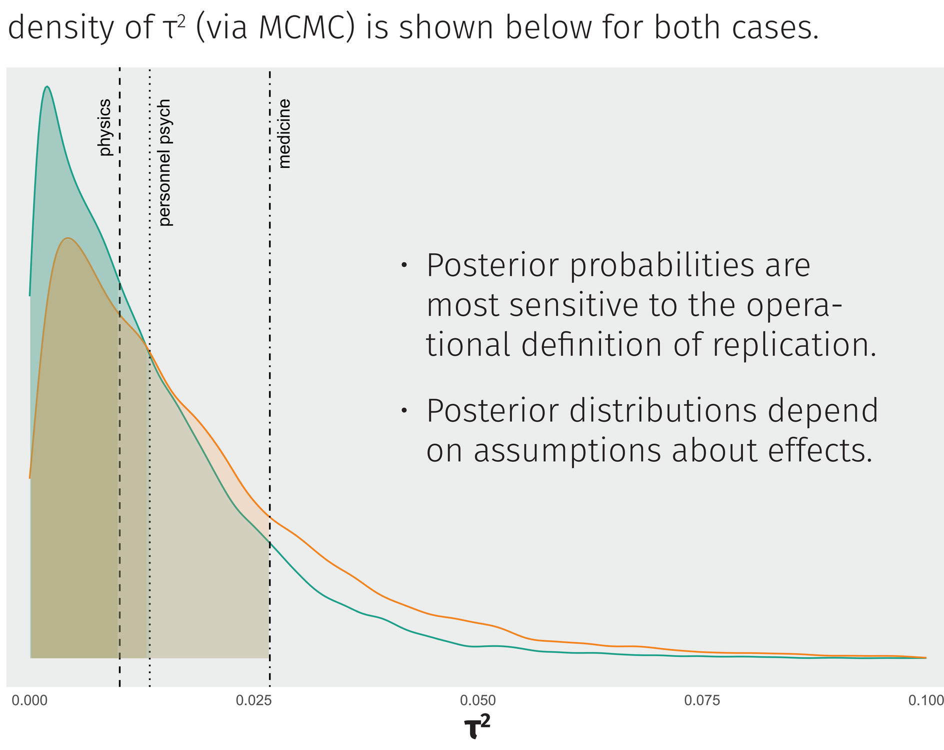
We use data from the ManyLabs Gambler's Fallacy experiments to illustrate these methods (Klein et al, 2014). The data comprise 36 standardized effect size estimates *g* (Hedges & Olkin, 1985).

- Heterogeneity statistic: Q = 51.609
- Precision-weighted mean: T. = 0.627
- Average within-study variance: v = 0.040

We use two models for the θ_i (normal and t-distributed with 4 df) and three priors for τ. We examine posterior probabilities of approximate replication, and posterior predictive checks for the heterogeneity statistic Q.

RESULTS: POSTERIOR INFERENCE p(τ² | T)

While our conclusions are less sensitive to our prior beliefs, they are sensitive to the operational definition of 'approximate replication', and our assumption of whether the effect parameters θ_i arose from a normal distribution or from a t-distribution with 4 degrees of freedom. The posterior density of τ^2 (via MCMC) is shown below for both cases.



Normal distribution

Given the data, the probability of approximate replication is 89%.

- $P[\tau^2 \le v/4 \mid \text{Data}] = 0.51$
- $P[\tau^2 \le v/3 \mid \text{Data}] = 0.62$
- $P[\tau^2 \le 2v/3 \mid \text{Data}] = 0.88$
- CI95(τ^2): [0.0001, 0.044]

t-distribution

Given the data, the probability of approximate replication is 79%.

- $P[\tau^2 \le v/4 \mid \text{Data}] = 0.39$
- $P[\tau^2 \le v/3 \mid \text{Data}] = 0.50$
- $P[\tau^2 \le 2v/3 \mid \text{Data}] = 0.78$
- CI95(τ^2) = [0.001, 0.059]

CONCLUSIONS & FUTURE WORK

- Assessments of replication should account for the fact that most replicates are approximate rather than exact, and address replication discrepancies rather than statistical issues. One way to do this is to model the effect parameters as if they were drawn at random from a distribution, and analyze their heterogeneity.
- The Bayesian framework offers a way to quantify our conclusions and check assumptions about the likelihood of replication.
- *Ultimate conclusions about whether studies replicate depend largely on the operational definition of replication. No conventional definition exists in education science.*
- Future efforts can leverage empirical evidence of heterogeneous effects in multi-site trials, as well as expert consensus to determine potential definitions of replication in education.

ASSESSING REPLICATION

Since experiments in education often sample from different populations and treatment fidelity can fluctuate, it would seem inevitable that study results of even successful replications might vary slightly. Future replication attempts in education science require practical definitions of replication that incorporate notions of "approximate" replication. We demonstrate how we might operationalize this definition by:

1. defining how we might quantify studies getting 'almost the same' results,
2. using Bayesian estimation to assess whether studies approximately replicate, and
3. using this method on an example from the ManyLabs replication project.

BAYESIAN INFERENCE

Bayesian models allow us to make explicit claims regarding our uncertainty about replication given the observed results.

PRIOR BELIEFS p(τ)

We quantify our a prior beliefs about replication using a prior distribution that describes our uncertainty about τ^2. In general, we recommend a prior distribution that reflects prior ignorance about τ^2. Some common choices include:

- $p(\tau) \sim \text{Uniform}(0, N)$: τ may be any value from 0 to N.
- $p(\tau) \sim$ Cauchy
- $p(\tau) \sim$ Inverse-Gamma

Note that we can also set our prior beliefs of exact replication with the prior: $p(\tau=0) = 1$.

PROBABILITY MODEL p(T | θ, τ²)

The probability model generates the data we observe:

- $\theta_i \sim G(\mu, \tau^2)$: Draw θ_i from a distribution G characterized by τ^2. Two possible choices include:
  - Normal: $\theta_i \sim N(\mu, \tau^2)$
  - t-distribution: $\theta_i \sim t(4, \mu, \tau^2/2)$
- $T_i \mid \theta_i \sim N(\theta_i, v_i)$: T_i has sampling variance v_i.

POSTERIOR INFERENCE P(τ² | T)

We update our beliefs in light of the data (T_i) that we observe via the posterior distribution of τ^2. We can use the posterior distribution to make probabilistic statements about the likelihood that a finding replicates given the studies observed. **Our assessment: $P(\tau^2 \le 2v/3 \mid T)$ = probability of replication.**

PREDICTIVE CHECKS p(Q_rep | T)

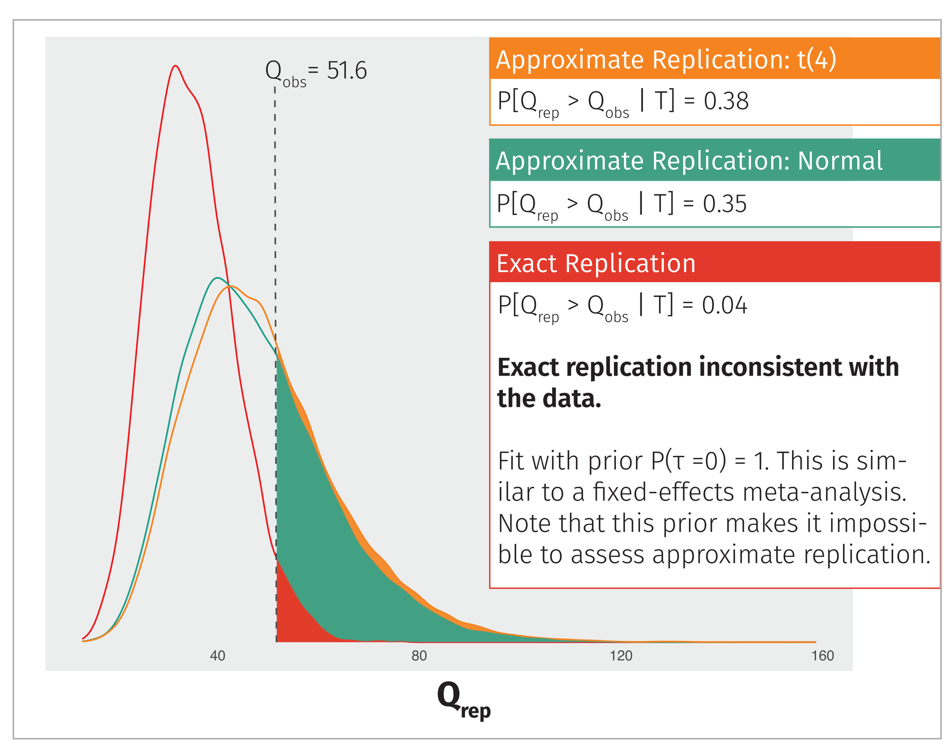
We can check models with the posterior predictive distribution. Here we use the heterogeneity statistic

$$Q = \Sigma(T_i - T.)^2/v_i$$

We compare the observed Q_{obs} to its predictive distribution given our updated beliefs about the parameters.

RESULTS: POSTERIOR PREDICTION

We can check models for exact and approximate replication using the posterior predictive distribution of the heterogeneity statistic Q. If our model is reasonable then the observed $Q_{obs} = 51.6$ should not fall near the extremes of the posterior predictive distribution $p(Q_{rep} \mid T)$. This occurs for the approximate replication models, but it does not for an exact replication model.



REFERENCES

- Hedges, L. V. (1987). How hard is hard science, how soft is soft science?: The empirical cumulativeness of research. American Psychologist, 42, 443-455.
- Hedges, L. V. & Olkin, I. (1985). Statistical Methods for Meta-Analysis. New York: Academic Press.
- Hedges, L. V. & Vevea, J. L. (1998). Fixed and random effects models in meta-analysis. Psychological Methods, 3, 486-504.
- Hedges, L. V. & Pigott, T. D. (2001). The power of statistical tests in meta-analysis. Psychological Methods, 6(3), 203-217.
- Klein, R. A. et al. (2014). Investigating variation in replicability: A "many labs" replication project. Social Psychology, 45(3), 142-152. doi:10.1027/1864-9335/a000178.
- Open Science Collaborative (2016). Estimating the reproducibility of psychological science. Science, 349, 943-951.