THE REPLICATION CRISIS

Replicating scientific findings raises important questions about the body of knowledge in education research. That span different fields—medicine, psychology, social sciences, etc.—and is often influenced by factors such as the quality of the research design, sample size, and statistical power. The replication crisis highlights the need for rigorous and transparent research practices.

PARAMETERS OF REPLICATION

For each replication, we have a true effect size $\theta_i$ that is estimated by $\hat{\theta}_i$, with sampling variance $v_i$. Inferences about $\theta_i$ are based on the distribution of $\theta_i$ around the true treatment effect $\theta_i$. Here, we characterize the $\theta_i$ as being Normally distributed around the true treatment effect $\theta_i$, with variance $v_i$.

PARAMETERS OF REPLICATION

Effect Parameters

Without knowing how the true effect parameters $\theta_i$ 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 generated $\theta_i$ that are very similar, we would conclude that the replicable findings are inferences about replication given the distribution from which the $\theta_i$ are drawn.

Heterogeneity

Differences among the study results can be characterized by $\tau^2$, the variance of the distribution that generated them. Our conclusions about replication will depend on the magnitude of $\tau^2$. How large a value of $\tau^2$ corresponds to approximate replication is a matter of scientific judgment. Here we consider $\tau^2$ relative to the observed sampling variances $v_i$, and the use of various estimates of $\tau^2$. We recommend a prior distribution that reflects prior ignorance on $\tau^2$, i.e., is noninformative.

DATA GENERATING PROCESS

The probability model generates the data we observe: $\theta_i$ and $\tau^2$. We use two models for $\theta_i$ (normal and t-distributed with 4 df) and three priors for $\tau^2$. We examine posterior probabilities of approximate replication, and posterior predictive checks for the heterogeneity statistic $Q$.

RESULTS: POSTERIOR INFERENCES

The posterior distributions depend on prior beliefs and the observed data. The posterior distribution of $\tau^2$ is shown below for both cases.

REFERENCES


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 from a random 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. Conventional definitions exist in education science.

ASSESSING REPLICATION

Since experiments in education often sample from different populations and treatments, the true effect might 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:
- defining how we might quantify studies getting 'almost the same' results,
- assessing whether studies approximately replicate, and
- using this method on an example from the ManyLabs replication project.