Replication in Education Research: Bayesian Meta-Analytic Perspectives

Jacob Schauer, Northwestern University

Background

The idea that findings replicate is fundamental to scientific progress. However, recent empirical analyses have cast doubt on the replicability of studies in various fields, including psychology and economics (Camerer et al., 2016; Ebersole et al., 2016; Ioannidis, 2005; Klein et al., 2014). This has major implications for IES; how can we hope to improve education if we cannot distinguish actual findings from anomalies? It is no surprise that the education science community has started to take this threat seriously, which involves supporting and publishing empirical evaluations of replication (Hedges, 2017).

While this work is important, the way in which we design and analyze these evaluations requires greater attention. In general, there are a variety of possible definitions of replication, and even more ways to measure it (Bollen et al., 2015; Lindsay, 2015). The modal approach—which we focus on for this poster—has been to conduct *direct replications* that seek to re-create an experiment as precisely as possible (S. Schmidt, 2009). Efforts in the social sciences have used a patchwork of metrics to analyze replication studies, and not all of them exhibit useful statistical properties (Valentine et al., 2011). Moreover, absent a clear analysis method, programs of research in replication are not designed to ensure that their inferences are appropriately conclusive. In this poster, we present a meta-analytic approach to defining and assessing replication and demonstrate it using data from the Many Labs replication project in psychology (Klein et al., 2014).

Methods

It has proven somewhat difficult to determine whether direct replications have successfully recreated a finding (S. Schmidt, 2009). The meta-analytic perspective casts this problem in terms of differences between underlying effect parameters, rather than of estimates or p-values (see Hedges & Olkin, 1985). Concretely, suppose there are $k \geq 2$ replicate studies. These studies involve effect parameters $\theta_i, i = 1, \ldots, k$, which reflect a sample from some distribution with unknown mean μ and variance τ^2. The θ_i may vary as a result of differences in sample compositions or experimental contexts across studies. We do not actually observe these parameters, but we do get an estimate T_i of θ_i. A common assumption in meta-analysis is that T_i are normally distributed with mean θ_i and known variance v_i.

We can define replication entirely in terms of the θ_i; if they are similar, then we might conclude the studies successfully replicate. A natural metric for their similarity is their unexplained variation τ^2. However, this definition requires we decide the proper magnitude of τ^2 that corresponds with replication. For example, we may consider replication as occurring when all of the θ_i are equal (and hence $\tau^2 = 0$). However, this may be too strict to be useful in practice; even in the hard sciences like particle physics, there is an expectation that experimental results may vary slightly (Hedges, 1987; Olive, 2014). It may be more reasonable to define some threshold τ_0^2 such that $\tau^2 \leq \tau_0^2$ would characterize negligible differences between studies. The proper value of τ_0^2 is a matter of scientific judgement, and may depend on the conventions in a given field. No such convention exists in education science, and must be part of future work. However, we examine how different definitions of replication might be operationalized using potentially relevant conventions from other fields—medicine, personnel psychology, and physics (see Hedges & Pigott, 2001).

While there are many ways to estimate τ^2, we use a Bayesian hierarchical model. This allows us to draw conclusions about the probability of approximate replication given the data. Concretely, the model is:

$$\mu \sim p(\mu)$$
$$\tau^2 \sim p(\tau^2)$$
$$\theta|\mu, \tau^2 \sim p(\theta|\mu, \tau^2)$$
$$T|\theta, v \sim N(\theta, v)$$

We use a few different choices for the distributions above to assess sensitivity (for further discussion, see Gelman, 2014, especially ch. 5.7). We model $p(\tau^2)$ as an inverse-gamma, Cauchy, and uniform distribution; and $p(\theta|\mu, \tau^2)$ as a normal distribution and a t-distribution with four degrees of freedom:

- $p(\mu)$ noninformative
- $p(\tau^2) \sim IG(0.1, 0.1)$, $p(\tau^2) \sim Cauchy(0, 2)$, and $p(\tau^2) \sim Uniform(0, 10^8)$
- $p(\theta|\mu, \tau^2) \sim N(\mu, \tau^2)$ and $p(\theta|\mu, \tau^2) \sim t(\mu, \tau^2, 4)$

Data

The data comprise 36 standardized mean differences and associated sampling variances from a Many Labs replication experiment: the "Reverse Gambler's Fallacy" (Klein et al., 2014). In the studies, participants were randomly assigned to one of two conditions and asked to imagine a man rolling dice at a casino. In one condition, subjects imagined seeing the man roll three sixes. In the other, they imagined him rolling two sixes and a three. Subjects were then asked how many times they thought the man had rolled the dice before they witnessed the result in their assigned condition. On average, participants who imagined seeing three sixes tended to estimate the man had rolled the dice more times than those who imagined seeing only two sixes. This poster uses data from the Many Labs replicates, but does not include the original finding by Oppenheimer & Monin (2009) for reasons related to publication bias (for discussion, see Hedges & Vevea, 1996).

Results

Given the data, the probability of approximate replication $p(\tau^2 \leq \tau_0^2|T)$ depends mainly on the choice of τ_0^2, as well as $p(\theta|\mu, \tau^2)$. If we assume the θ are normally distributed, then the probability that the replications were successful given the data range from 0.89—if we consider the least stringent definition of replication (largest τ_0^2)—to 0.53—for the most stringent convention (smallest τ_0^2). However, if the θ follow a location-scale t-distribution with four degrees of freedom—which has much wider tails—then the probability that the studies replicate is lower, ranging from 0.79 to 0.42, depending on the operational definition of replication (τ_0^2). What this illustrates is that analyses of replicate studies require careful consideration of (a) a precise and well-justified definition of replication, and (b) the actual distribution of the study results.

In general, we advocate for using both the meta-analytic model and the concept of approximate replication. The former brings a principled and well-studied approach to important questions about our ability to reproduce findings, while the latter provides analyses that are practical and meaningful. Moreover, while it is not necessary to use the Bayesian formalism for this framework, it does provide an intuitive way to describe conclusions about replication. That said, more work must be done to determine a useful operational definition of replication in the field of education science.

References

Bollen, K., Cacioppo, J. T., Kaplan, R. M., Krosnick, J. A., & Olds, J. L. (2015). *Social, behavioral, and economic sciences perspectives on robust and reliable science* (Report of the Subcommittee on Replicability in Science Advisory Committee to the National Science Foundation Directorate for Social, Behavioral, and Economic Sciences). National Science Foundation. Retrieved from https://www.nsf.gov/sbe/AC_Materials/SBE_Robust_and_Reliable_Research_Report.pdf

Camerer, C. F., Dreber, A., Forsell, E., Ho, T.-H., Huber, J., Johannesson, M., ... Wu, H. (2016). Evaluating replicability of laboratory experiments in economics. *Science*, *351*(6280), 1433–1436. doi:10.1126/science.aaf0918

Ebersole, C. R., Atherton, O. E., Belanger, A. L., Skulborstad, H. M., Allen, J. M., Banks, J. B., ... Nosek, B. A. (2016). Many labs 3: Evaluating participant pool quality across the academic semester via replication. *Journal of Experimental Social Psychology*, *67*, 68–82. doi:10.1016/j.jesp.2015.10.012

Gelman, A. (2014). *Bayesian data analysis* (Third edition). Boca Raton: CRC Press.

Hedges, L. V. (1987). How hard is hard science, how soft is soft science? The empirical cumulativeness of research. *American Psychologist*, *42*(5), 443–455. doi:10.1037/0003-066X.42.5.443

Hedges, L. V. (2017). Challenges in building usable knowledge in education. *Journal of Research on Educational Effectiveness*, 00–00. doi:10.1080/19345747.2017.1375583

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis.* Orlando: Academic Press.

Hedges, L. V., & Pigott, T. D. (2001). The power of statistical tests in meta-analysis. *Psychological Methods*, *6*(3), 203–217. doi:10.1037//1082-989X.6.3.203

Hedges, L. V., & Vevea, J. L. (1996). Estimating effect size under publication bias: Small sample properties and robustness of a random effects selection model. *Journal of Educational and Behavioral Statistics*, *21*(4), 299–332. doi:10.3102/10769986021004299

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine*, *2*(8), e124. doi:10.1371/journal.pmed.0020124

Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B., Bahník, Š., Bernstein, M. J., ... Nosek, B. A. (2014). Investigating variation in replicability: A "many labs" replication project. *Social Psychology*, *45*(3), 142–152. doi:10.1027/1864-9335/a000178

Lindsay, D. S. (2015). Replication in psychological science. *Psychological Science*, *26*(12), 1827–1832. doi:10.1177/0956797615616374

Olive, K. A. (2014). Review of particle physics. *Chinese Physics C*, *38*(9), 090001.

Oppenheimer, D. M., & Monin, B. (2009). Investigations in spontaneous discounting. *Memory & Cognition*, *37*(5), 608–614. doi:10.3758/MC.37.5.608

Schmidt, S. (2009). Shall we really do it again? The powerful concept of replication is neglected in the social sciences. *Review of General Psychology*, *13*(2), 90–100. doi:10.1037/a0015108

Valentine, J. C., Biglan, A., Boruch, R. F., Castro, F. G., Collins, L. M., Flay, B. R., ... Schinke, S. P. (2011). Replication in prevention science. *Prevention Science*, *12*(2), 103–117. doi:10.1007/s11121-011-0217-6

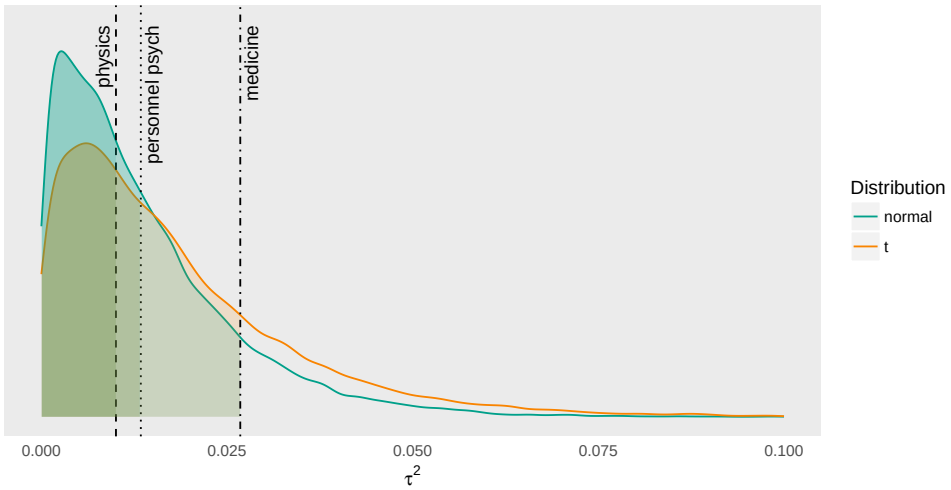

Figure 1: This figure displays the posterior distribution of $\tau^2|T$ for the case where θ are normally distributed (green) and t-distributed with four degrees of freedom (orange). The dashed lines are drawn at $\tau^2 = \bar{v}/4$, $\tau^2 = \bar{v}/3$, and $\tau^2 = 2\bar{v}/3$, which reflect conventions of negligible heterogeneity from physics, personnel psychology, and medicine, respectively. For each, these values correspond to the largest τ^2 may be to still conclude that the studies approximately replicate. The actual posterior probability of replication depends strongly on which convention is deemed appropriate.