Poster #1
READS for Summer Learning: A Case Study on Systematic Replication

Background

In recent years, there has been renewed interest in replication in the education sciences as evidence by replication reform efforts (e.g. Spybrook et al., 2019), an increase in methodological research (e.g. Hedges 7 Schauer, 2018; Wong & Steiner, 2018), and examinations of funding structures for replication (e.g. Chhin et al., 2018). In response to this work, IES recently established a grant competition to fund replication programs that will study one of thirteen pre-selected interventions in "carefully chosen venues" that *systematically* vary in context, methods, implementation, outcomes, and intervention components (IES RFA, 2019). Although systematic replication has been considered on a philosophical level (Sidman, 1960; Lykken, 1968; Hendrick, 1991), little is understood about what it might look like in practice including what the goals of systematic replication might be and how a funding agency (or primary investigator) might design a replication program to achieve those goals.

Purpose

The goal of this study is to better understand how experimental designs have been used to account for systematic variation in existing replication programs and what limitations these designs may create in assessing the goals of a replication program. I use a case study to identify goals for a replication program, define statistical tests relevant to each goal, and demonstrate how the experimental design of the contributing studies can lead to issues of confounding when assessing these tests.

Methods

***Sample*.** In order to find an appropriate case study, I completed a search of U.S. Department of Education funded grants for replication from 2004 to 2018 (e.g. IES, i3). I reviewed studies explicitly labeled as replications as well as follow-up, scale-up, and effectiveness studies. I defined a replication as any study that (1) referred to an initial efficacy study in the grant summary and (2) attempted to study the same phenomenon or intervention using newly collected data. Once a replication study was identified, I completed a Google search of the intervention in question to identify any additional replication studies that might have been funded by another agency. This search uncovered a handful of replication programs that consist of more than one replication study. For the purpose of this study, I will limit consideration to the most comprehensive of these programs: READS for Summer Learning.

READS for Summer Learning is a voluntary summer reading interventions consisting of one pilot study, an initial efficacy study, and eight subsequent replications, each of which varies one or more parameters of the initial studies including geographical location, sample demographics, treatment components, data collection tools and procedures, and implementation.

Data Collection. In order to understand systematic variation in the context of this program, I extracted data from each study to build three datasets. The first dataset contains information on the following dimensions:

- Study design (e.g. experimental design, number of treatment conditions)
- Sample characteristics (e.g. urbanicity, percent minority)
- Treatment components (e.g. book dosage, family literacy events)
- Observing operations (e.g. reading achievement measures)

The second dataset contains research questions, effect sizes, and standard errors for each study in the program. The final dataset contains rhetoric from each publication about the individual study's goals and conclusions about those goals.

Results

Goals of Systematic Replication. The data reveal that assessment of systematic replication may involve assessing one or more program goals:

1. *Replication*, in which the authors test the similarity in effect parameters under similar conditions
2. *Generalizability*, in which the authors test the similarity in effect parameters under different conditions (e.g. new populations)
3. *Refinement,* in which the authors assess the impact of alterations to existing components or addition of new components improves the overall impact of the intervention
4. *Fidelity/Precursors-to-Scale,* in which the authors assess whether the intervention could be feasibly implemented in natural conditions

The authors tested for replication and generalizability by comparing the direction and /or significance of pairs of effect size estimates. It has since been shown that these metrics are flawed and using a meta-analytic approach is preferable (see Schauer, under review). Thus, I propose considering tests for replication and generalizability of the form $H_0: \lambda \leq \lambda_0$ where $\lambda = \sum \frac{(\theta_i - \theta_\cdot)^2}{v_i}$ for effects $\theta_1, \dots, \theta_k$ and λ_0 is a negligible value of heterogeneity (see Hedges & Schauer, 2018). Tests for refinement come in two forms. In some studies, the authors employ experimental designs with more than two treatment arms such that internal comparison of effects can be used to assess refinement. When testing for refinement using a set of two-arm RCTs, however, the authors must rely on external comparisons of effects. That is, they test $H_0: \theta_j \leq \max\{\theta_1, \dots, \theta_{j-1}\}\ vs.\ H_A: \theta_j > \max\{\theta_1, \dots\ \theta_{j-1}\}$. Interestingly, should a refinement be successful, studies that did not include the refinement are no longer included in the tests for replication.

Understanding Confounding Structures. I use six theoretically important intervention components studies over the course of READS and, for simplicity, coded each to have two levels. From these six factors, I create a confounding structure for the entire program and situate each of the READS studies within this structure. Based on this approach, there are 64 possible treatment conditions in the READS program, 6 of which are actually explored by READS. In each study, the effects of interest in any of the tests mentioned above are aliased with at least 15 other effects, whether in a positive or negative direction. While some of these effects are higher-order interactions that may not be theoretically important, others are main effects or two-way interactions that are likely to be important.

Conclusion

In the analyses reported in this proposal, I found that the goals of systematic replication are not as straightforward as they might appear. Further, using two-arm RCTs to explore systematic variation and draw conclusions about replication can result in some serious issues of confounding. I will use these analyses as a starting point for discussions about how we might design replication programs to address the various goals of systematic replication and to minimize confounding between theoretically important study parameters.

References

Chin, C.S., Taylor, K.A., & Wei, W.S. (2018). Supporting a Culture of Replication: An Examination of Education and Special Education Research Grants Funded by the Institute of Education Sciences. *Educational Researcher, 47*(9), 594-605.

Hedges, L.V., & Schauer, J.M. (2018). Statistical Analyses for Study Replication: Meta-analytic Perspectives. *Psychological Methods.*

Hendrick, C. (1990). Replications, strict replications, and conceptual replications: are they important?. *Journal of Social Behavior and Personality, 5*(4), 41.

Institute of Education Sciences (2019). Request for applications: Research grants focused on systematic replication. Retrieved from http://ies.ed.gov/funding/pdf/2020_84305R.pdf.

Lykken, D.T. (1968). Statistical significance in psychological research. *Psychological bulletin, 70*(3p1), 151.

Sidman, M. (1960). Tactics of Scientific Research.

Spybrook, J., Anderson, D., & Maynard, R. (2019). The Registry of Efficacy and Effectiveness Studies (REES): A Step Toward Increased Transparency in Education. *Journal of Research on Educational Effectiveness, 12*(1), 5-9.

Wong, V.C. & Steiner P.M. (2018). *Replication Designs for Causal Inference*. EdPolicyWorks Working Paper Series No. 62. Available at http://curry.virginia/edu/edpolicyworks/wp.