Symposium: The Methodological Foundations of Replication Sciences

Motivation

Efforts to promote evidence-based practices in education (e.g., What Works Clearinghouse) assume that scientific findings are of sufficient validity to warrant its use in decision making. Replication has long been a cornerstone for establishing trustworthy scientific results. At its core is the belief that scientific knowledge should not be based on chance occurrences. Rather, it should established through systematic and transparent methods, results that can be independently replicated, and findings that are generalizable to at least some target population of interest (Bollen, Cacioppo, Kaplan, Krosnick, & Olds, 2015).

Given the central role of replication in the accumulation of scientific knowledge, researchers have reevaluated the replicability of seemingly well-established findings. Results from these efforts have not been promising. The Open Science Collaboration (OSC, 2105) replicated 100 experimental and correlational studies published in high impact psychology journals. Overall, the OSC found that only 36% of these efforts produced results with the same statistical significance pattern as the original study. Ioannidis (2008) reviewed more than 1,000 medical publications and found that only 44% of replication efforts produced results that corresponded with the original findings. Combined, these results contribute to a growing sense of a "replication crisis" occurring in multiple domains of science, including education (Makel & Plucker, 2014), prevention science (Valentine et al., 2011), and economics (Duvendack, Palmer-Jones, & Reed, 2017).

Content & Format of Symposium

Despite consensus on the need to promote replication efforts, there remains considerable disagreement about what constitutes as replication, how a replication study should be implemented, how results from these studies should be interpreted, and whether direct replication of results is even possible. The proposed symposium addresses these concerns by presenting multiple perspectives on replication, as well as by discussing how replication designs may be used and analyzed to assess the replicability and generalizability of effects and to identify treatment effect heterogeneity.

The symposium will include three paper presentations and commentaries provided by two expert discussants. **Vivian C. Wong & Peter M. Steiner** will introduce the Causal Replication Framework, which formalizes conditions under which replication success can be expected. The framework identifies all the assumptions needed for the direct replication of results and shows how different replication designs can be derived from this framework and used to evaluate treatment effect heterogeneity. **Jacob Schauer** will frame replication from a meta-analytic perspective and discuss tests for assessing replication success. The definition of replication corresponds with identifying heterogeneity between underlying parameters, and requires careful consideration of how studies are being conducted and how similar we can expect results to be. Finally, **Bryan Keller** will present the theory and an example of a novel six-arm replication design (i.e., within-study comparison) for examining whether propensity score and regression adjustments to an observational study with self-selection can replicate the results of a corresponding randomized experiment.

Two well-known statisticians and experts in replication and causal generalization, **Larry Hedges** and **Elizabeth Stuart**, will discuss the three proposed papers and provide more general commentaries replication issues.

A Causal Replication Framework for Designing and Assessing Replication Efforts

Vivian C. Wong (University of Virginia—Charlottesville)
Peter M. Steiner (University of Wisconsin—Madison)

Background

Replication has long been a cornerstone for establishing trustworthy scientific results. At its core is the belief that scientific knowledge should not be based on chance occurrences. Rather, it should be established through systematic and transparent methods, results that can be independently replicated, and findings that are generalizable to at least some target population of interest (Bollen, Cacioppo, Kaplan, Krosnick, & Olds, 2015; Zwaan et al., 2018). Despite consensus on the need to promote replication efforts, there remains considerable disagreement about what constitutes a replication, how a replication study should be implemented, how results from these studies should be interpreted, and whether direct replication of results is even possible.

Objectives / Research Questions

We address these concerns by presenting the methodological foundations for a "replication science." The paper introduces the *Causal Replication Framework*, which uses potential outcomes notation to formalize the conditions under which a successful replication of a *causal effect* can be expected. The Causal Replication Framework begins with the premise that replication is for "*establishing a fact, truth, or piece of knowledge*" (Schmidt, 2009). Here, the "piece of knowledge" can be described as the causal effect of a well-defined treatment-control contrast on a clearly specified outcome of an explicitly defined target population. We refer to this unknown causal effect as the *causal estimand* which is the target of inference in the original and replication study. Under the Causal Replication Framework, the goal is to identify and estimate the same well-defined causal estimand of interest in both studies. When results do not replicate, it is because one or more replication design assumptions has not been met.

The Causal Replication Framework differs from most current definitions of replication, which focus on the replication of methods and procedures (e.g., Brandt et al., 2014; Nosek and Errington, 2017; OSC, 2015; see also Schmidt, 2009). Under the Causal Replication Framework, while two studies *may* use the same procedures and methods to generate the same corresponding causal effect, it is also possible for studies to use different methods and procedures *as long as they identify and estimate the same well-defined causal estimand of interest*. The focus on replication design assumptions (instead of methods and procedures) has the potential to help researchers understand how prospective and post-hoc replication designs can be used to systematically assess the replicability of effects, as well as to evaluate potential sources of treatment effect heterogeneity when study results do not replicate. Thus, the Causal Replication Framework may be understood as "causal" in two ways – first, its assumptions are focused on the replicability of *causa*l effects; second, it shows how research designs may be used to identify *why* results do not replicate in a causal way.

Method / Findings / Application

Under the Causal Replication Framework (Wong & Steiner, 2018), successful replication of effects (within the limits of sampling error) can be expected if the causal estimands in both the original and replication study are identical. This means that the causal estimand of interest must

be identifiable and estimable without bias in each study. Five replication design assumptions (formalized using potential outcomes) must be met for two studies to produce the same effect. They include:

A1 Treatment & Outcome Stability
A2 Equivalence of Causal Estimands
A3 Identification of Causal Estimands
A4 Unbiased Estimation of Causal Estimands
A5 Estimands, Estimators, and Estimates are Correctly Reported

Conceptualizing replication through the Causal Replication Framework yields two important implications for practice. First, although assumptions for the direct replication of results are stringent in practice, we show that it is possible for researchers to address and probe these assumptions through the thoughtful use of research designs and empirical diagnostic tests. Second, the framework may be used to derive replication designs for "conceptual replication" approaches. Such replication designs follow in cases where researchers decide to deliberately violate one or more replication assumptions, such as introducing systematic variation across studies in population and setting characteristics, treatment and control conditions, and/or research methods for identifying and estimating effects (as done by within-study comparisons). When replication design assumptions are systematically tested across study, it is possible to identify sources of treatment effect heterogeneity.

To highlight the practical usefulness of the Causal Replication Framework, we will discuss the strength and weaknesses of some replication designs through two empirical examples. We will demonstrate that research designs for replication are both feasible and desirable for producing usable, robust causal knowledge.

Conclusions / Implications

The Causal Replication Framework allows for a clear formulation of the assumptions required to successfully replicate a causal effect in two studies. The main advantages of this assumption-based replication framework are that researchers can (a) formulate and discuss replication efforts along clearly formulated causal replication assumptions, (b) think about design elements and tests to empirically defend these assumptions, and (c) derive a broad range of replication design variants by systematically relaxing one or multiple assumptions. The paper will demonstrate how replication design assumptions may be addressed and probed empirically through two empirical examples.

References

Bollen, K., Cacioppo, J., Kaplan, R., 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*, 1–29.

Brandt, M. J., et al. (2014). The Replication Recipe: What makes for a convincing replication? *Journal of Experimental Social Psychology*, *50*(1), 217–224.

Nosek, B. A., & Errington, T. M. (2017). Making sense of replications. *ELife*, *6*.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251).

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.
Wong, V. C., & Steiner, P. M. (2018b). Replication Designs for Causal Inference. *EdPolicyWorks Working Paper Series*. Retrieved from http://curry.virginia.edu/uploads/epw/62_Replication_Designs.pdfhttp://curry.virginia.edu/edpolicyworks/wp

Assessing Replication: Lessons for Education Science

Jacob M. Schauer
Northwestern University

Background

Recent research has called into question the replicability of findings in various scientific fields, including psychology and behavioral economics (Open Science Collaboration, 2015). It would seem that it is only a matter of time before similar challenges arise in education science (Hedges, 2018). And while we might look to replication research programs in other social sciences for guidance, the methods used to analyze these programs have been subject to immediate and substantial criticism (D. T. Gilbert et al., 2016; J. C. Valentine et al., 2011).

Focus of Study

Given how important replication is to the logic and rhetoric of science, one would expect a standard approach to designing and analyzing replication studies. However, as has been noted by various researchers, this is not the case (S. Schmidt, 2009). Not only is there no standard analysis, it appears that there is not even a clear-cut definition of what it means for studies to successfully replicate (Bollen et al., 2015; Open Science Collaboration, 2015).

Recent research has argued that meta-analysis provides a framework to formalize definitions of replication and analyze replication studies (Hedges & Schauer, 2018). In this framework, a study's results are viewed as an underlying effect parameter, and replication would imply that parameters from different studies are similar in value. However, precisely how this is defined requires statistical and theoretical considerations, which will in turn affect analysis methods and their properties. This paper describes these considerations, outlines their statistical implications, and uses data from replication programs in the social sciences to shed light on how they might play out in practice.

Data

In the social sciences, there have been several high-profile replication programs, whose data are used in this paper. These programs include the Many Labs Replication Project, Pipeline Project (PPIR), APS's Registered Replication Reports (RRR), Replication Project: Psychology (RPP), and Replication Project: Economics (RPE); they are summarized in Table 1. All told, the data comprise results of replication studies of 132 different findings, each of which has been conducted between two and 37 times. Information regarding effect sizes was pulled from the Open Science Framework, often using the original investigators' own code.

Methods

Suppose that $k \geq 2$ replicate studies have been conducted. Denote θ_i as the underlying effect for study $i = 1, \ldots, k$. The parameter θ_i is what would be observed in the absence of any estimation error variance, such as from sampling. In practice, θ_i is estimated by T_i, which has some variance v_i. When effects are on

the scale of standardized mean differences (as in this paper) and studies have moderate to large sample sizes, estimates are unbiased and normally distributed:

$$T_i \sim N(\theta_i, v_i)$$

Replication can be conceived of in terms of how similar the θ_i are. However, precisely what this means depends on a few theoretical considerations. The first is how to parametrize differences between studies. If only two studies are conducted, then we might treat the two results θ_1 and θ_2 as fixed, but unknown constants, and we can define replication in terms of their difference $\theta_1 - \theta_2$. However, if several replications have been run, then one approach would be to model the studies and their effects using a random effects meta-analysis model (Hedges & Vevea, 1998). This assumes that the θ_i are a sample from a universe of possible effects, and inferences pertain to how similar effects are in that universe. One way parametrize replication in this context is with the variance τ^2 of the distribution from which the θ_i are a sample.

Another consideration is whether replication is exact or approximate. One possible definition is that the θ_i are identical, so that $\theta_1 = \ldots = \theta_k$ or $\tau^2 = 0$. However, replication could also occur if results that are "practically" the same but not identical. In that case, an idea of approximate replication would allow for the θ_i to be negligibly different or τ^2 to be negligibly small. Precisely what constitutes a small amount of heterogeneity among study results is subject to scientific judgement. In meta-anlaysis, different fields have communicated rules of thumb about negligible heterogeneity in the metric of τ^2/v_i that range from $\tau^2/v_i = 1/4$ (in physics, see Olive, 2014) to $2/3$ (in medicine, see Higgins & Green, 2008).

Hedges & Schauer (2018) propose hypothesis tests to incorporate these considerations, and argue that random effects meta-analysis methods may also be appropriate (i.e., for point estimates). Using this approach, this paper assesses the extent to which findings may have replicated in empirical research, and how sensitive these analyses are for the designs that have been used.

Results

There are two main results from data on replication programs that can be instructive for education science. The first is that we might expect *some* heterogeneity among pre-registered replications. Figure 1 shows boxplots of estimates of τ^2 for pre-registered replications by program. Note that replicates of many findings exhibit heterogeneity that is estimated to either zero, or to be on the order of what might be considered negligible in physics ($\tau^2/v = 1/4$) or psychology ($\tau^2/v = 2/3$). However, for several other findings, replications were considerably more heterogeneous.

The second is that the designs used so far have had limited sensitivity to differences between studies. Figures 2 and 3 show standard errors for estimated heterogeneity (on the scale of $\theta_1 - \theta_2$ or τ^2) by program. Figure 2 shows that designs that have used only $k = 2$ studies have been insensitive to potentially meaningful differences between studies. And while the Figure 3 shows that designs with $k \geq 2$ may be more sensitive, they are only sensitive to larger amounts of heterogeneity. Taken together, it would appear that further work is required on developing analyses of *approximate* replication, and designing appropriately sensitive ensembles of replication studies.

Program	**Paper**	**# Findings**	k
Replication Project: Psychology (RPP)	Open Science Collaboration, 2015	73	2
Replication Project: Economics (RPE)	Camerer et al., 2016	18	2
Registered Replication Reports (RRR)	Alogna et al., 2014	2	24–33
	Bouwmeester et al., 2017	1	22
	Cheung et al., 2016	4	17
	Eerland et al., 2016	3	13
	Hagger et al., 2016	2	24
	Wagenmakers et al., 2016	1	18
Many Labs Replication Project	Klein et al, 2014	17*	36–37
Pipeline Project (PIPR)	Schweinsberg et al., 2016	11	12–18
Total		132	12–37

TABLE 1: REPLICATION RESEARCH PROGRAMS IN THE SOCIAL SCIENCES. *This table summarizes the data used in this paper, which are results from replication studies conducted as part of different systematic programs of research on replication in the social sciences. For each program, the table displays the number of findings studied and how many times each study was replicated (k). Note that data from the RPP comprise its 'meta-analytic subset' of findings. For the Many Labs Project, it was determined that one experiment (on 'anchoring') involved two distinct experimental designs based on study results; one design took place in a lab, the other was conducted online. Thus, there are 17 findings (rather than the 16 reported in that paper) in this analysis.*

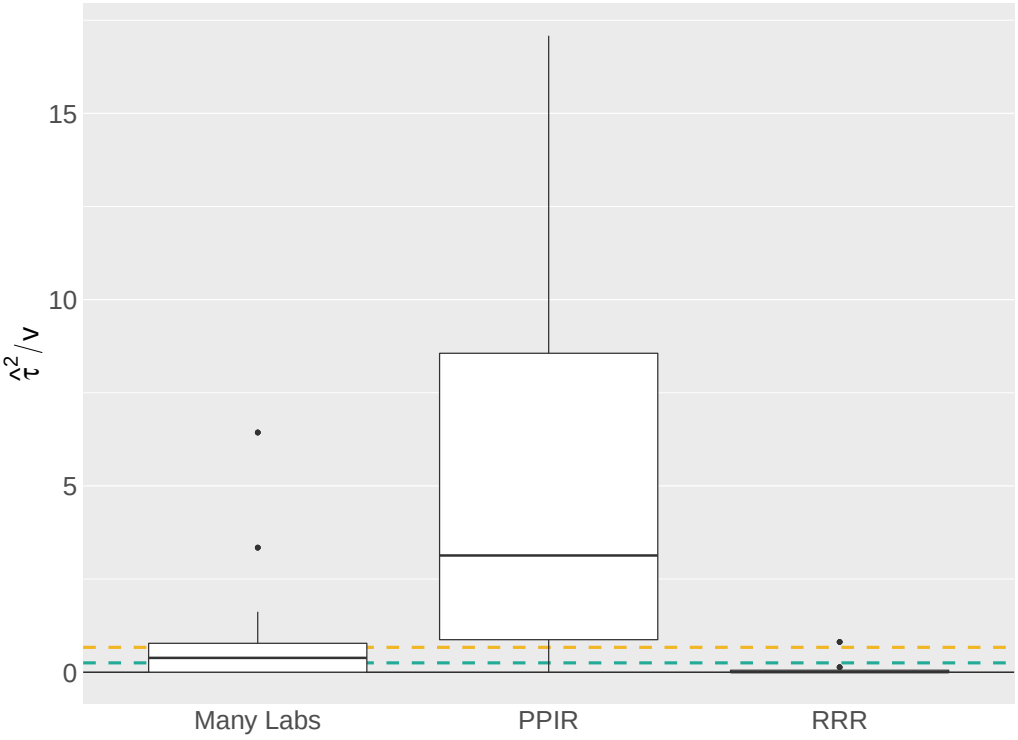

FIGURE 1: HETEROGENEITY OF MULTI-LAB REPLICATIONS. *This plot shows the distribution of estimated heterogeneity for three multi-lab replication programs. For each program on the x-axis, the boxplot shows point estimates of the variance of the study results, so that each data point of the boxplot represents an estimate for a given ensemble of replications for a single finding. Heterogeneity is displayed on the the scale of relative to the within-study variances (τ^2/v). Dashed lines correspond to values of heterogeneity considered negligible in meta-analyses in physics (green line $\tau^2/v = 1/4$) and personnel psychology (orange line $\tau^2/v = 2/3$).*

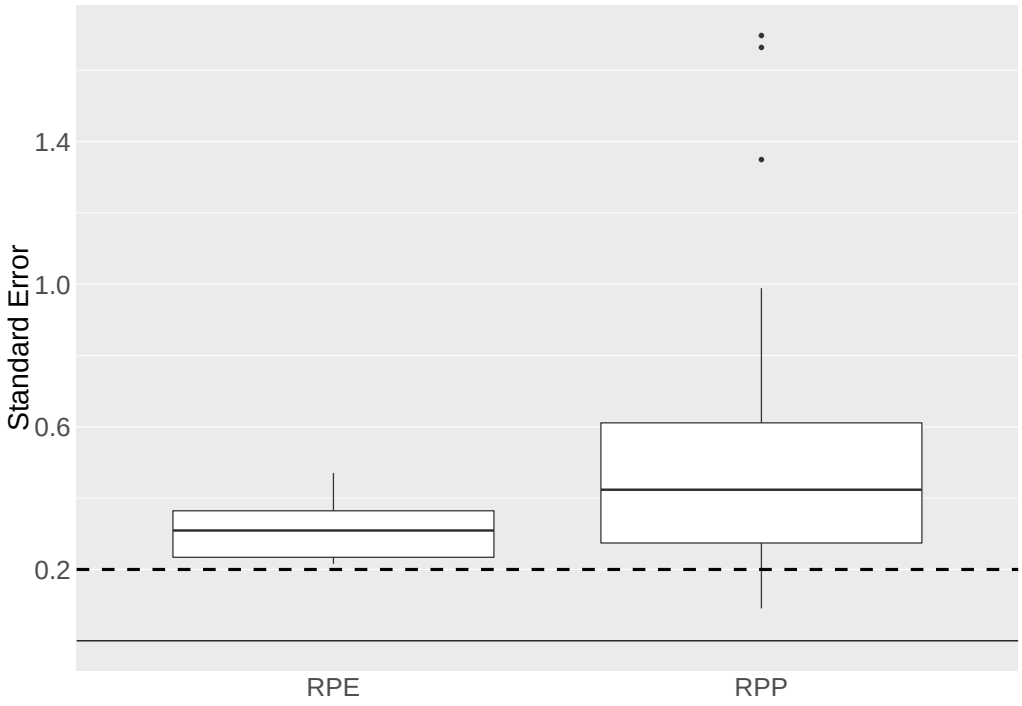

FIGURE 2: STANDARD ERROR OF ESTIMATED DIFFERENCES: RPP AND RPE. *This plot shows the distribution of standard errors for estimated differences between studies. For each program on the x-axis, the boxplot shows the standard error of the estimate of $\theta_1 - \theta_2$, so that each data point of the boxplot represents a standard error for a given pair of replications of the same experiment. The dashed line corresponds to a difference between effects of $|\theta_1 - \theta_2| = 0.2$ so that one study could have a small effect and the other would have zero effect. Since nearly all of the standard errors are greater than this, these designs would not have been sensitive to a difference in effects of 0.2.*

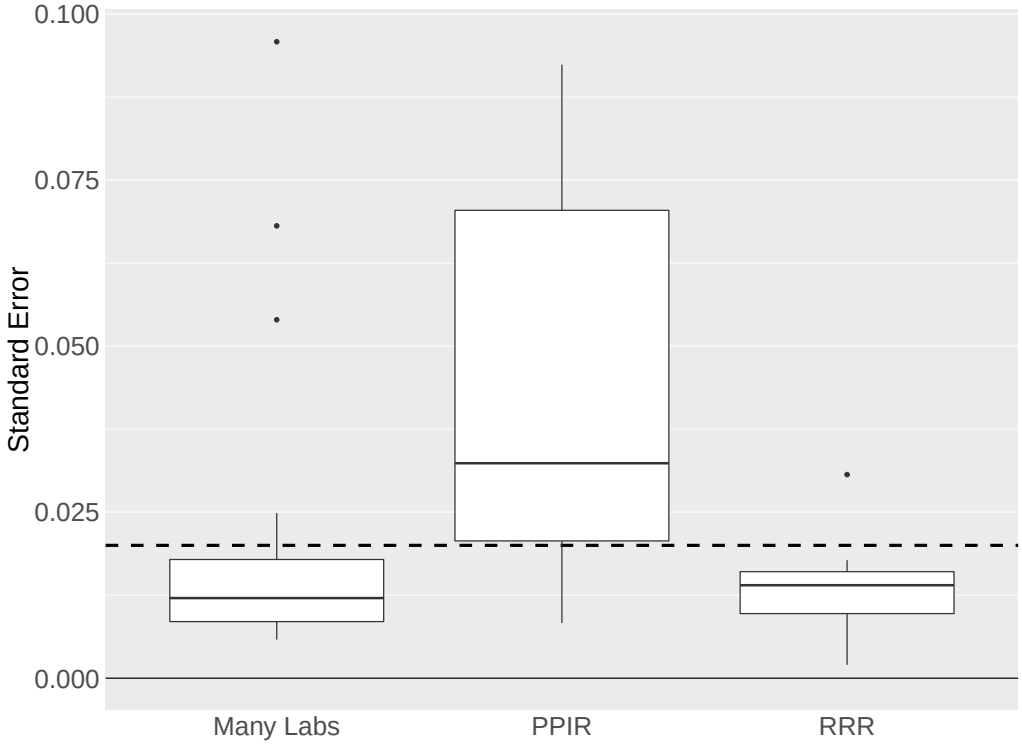

FIGURE 3: STANDARD ERRORS OF ESTIMATED HETEROGENEITY: MULTI-LAB REPLICATIONS. *This plot shows the standard error for estimated heterogeneity among multiple replication attempts. For each program on the x-axis, the boxplot shows the standard errors of the DerSimonian-Laird estimate of τ^2 for each finding. The dashed line corresponds to 0.02; heterogeneity of $\tau^2 = 0.02$ would correspond to most results being about the same size (i.e, within 0.15 of the mean effect in Cohen's d units). Note that some ensembles have standard errors smaller than this, though not by much, and that many ensembles had standard errors much larger.*

References

Alogna, V. K., Attaya, M. K., Aucoin, P., Bahník, Š., Birch, S., Birt, A. R., … Zwaan, R. A. (2014). Registered replication report: Schooler and Engstler-Schooler (1990). *Perspectives on Psychological Science*, *9*(5), 556–578. https://doi.org/10.1177/1745691614545653

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

Bouwmeester, S., Verkoeijen, P. P. J. L., Aczel, B., Barbosa, F., Bègue, L., Brañas-Garza, P., … Wollbrant, C. E. (2017). Registered replication report: Rand, Greene, and Nowak (2012). *Perspectives on Psychological Science*, *12*(3), 527–542. https://doi.org/10.1177/1745691617693624

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. https://doi.org/10.1126/science.aaf0918

Cheung, I., Campbell, L., LeBel, E. P., Ackerman, R. A., Aykutoğlu, B., Bahník, Š., … Yong, J. C. (2016). Registered replication report: Study 1 from Finkel, Rusbult, Kumashiro, & Hannon (2002). *Perspectives on Psychological Science*, *11*(5), 750–764. https://doi.org/10.1177/1745691616664694

Eerland, A., Sherrill, A. M., Magliano, J. P., Zwaan, R. A., Arnal, J. D., Aucoin, P., … Prenoveau, J. M. (2016). Registered replication report: Hart & Albarracín (2011). *Perspectives on Psychological Science*, *11*(1), 158–171. https://doi.org/10.1177/1745691615605826

Gilbert, D. T., King, G., Pettigrew, S., & Wilson, T. D. (2016). Comment on "estimating the reproducibility of psychological science". *Science*, *351*(6277), 1037–1037. https://doi.org/10.1126/science.aad7243

Hagger, M. S., Chatzisarantis, N. L. D., Alberts, H., Anggono, C. O., Batailler, C., Birt, A. R., … Zwienenberg, M. (2016). A multilab preregistered replication of the ego-depletion effect. *Perspectives on Psychological Science*, *11*(4), 546–573. https://doi.org/10.1177/1745691616652873

Hedges, L. V. (2018). Challenges in building usable knowledge in education. *Journal of Research on Educational Effectiveness*, *11*, 1–21. https://doi.org/10.1080/19345747.2017.1375583

Hedges, L. V., & Schauer, J. M. (2018). Statistical analyses for studying replication: Meta-analytic perspectives. *Psychological Methods*.

Hedges, L. V., & Vevea, J. L. (1998). Fixed- and random-effects models in meta-analysis. *Psychological Methods*, *3*(4), 486–504. https://doi.org/10.1037/1082-989X.3.4.486

Higgins, J. P. T., & Green, S. (Eds.). (2008). *Cochrane handbook for systematic reviews of interventions*. Chichester, England ; Hoboken, NJ: Wiley-Blackwell.

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. https://doi.org/10.1027/1864-9335/a000178

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

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716–aac4716. https://doi.org/10.1126/science.aac4716

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. https://doi.org/10.1037/a0015108

Schweinsberg, M., Madan, N., Vianello, M., Sommer, S. A., Jordan, J., Tierney, W., … Uhlmann, E. L. (2016). The pipeline project: Pre-publication independent replications of a single laboratory's research pipeline. *Journal of Experimental Social Psychology*, *66*, 55–67. https://doi.org/10.1016/j.jesp.2015.10.001

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. https://doi.org/10.1007/s11121-011-0217-6

Wagenmakers, E.-J., Beek, T., Dijkhoff, L., Gronau, Q. F., Acosta, A., Adams, R. B., … Zwaan, R. A. (2016). Registered Replication Report: Strack, Martin, & Stepper (1988). *Perspectives on Psychological Science*, *11*(6), 917–928. https://doi.org/10.1177/1745691616674458

A six-arm design replication study: Design, results, and implications

Bryan Keller

Teachers College, Columbia University

Author Note

A six-arm design replication study: Design, results, and implications

Background/Context

Design replication studies, also referred to as within-study comparisons, are broadly defined as studies that permit, by design, comparisons of effect size estimates from observational studies to estimates based on randomized experiments for the same intervention (Cook, Shadish, & Wong, 2008). Prior to 2008, design replication studies almost exclusively used a "three-arm" design, wherein the control group from a randomized experiment was replaced by a non-randomly selected comparison group; see Figure 1. Shadish, Clark, and Steiner (2008) proposed and implemented, and Pohl, Steiner, Eisermann, Soellner, and Cook (2009) replicated, a "four-arm" doubly-randomized preference design in which participants were randomly assigned to be in either a randomized experiment or a quasi-experiment; see Figure 2.

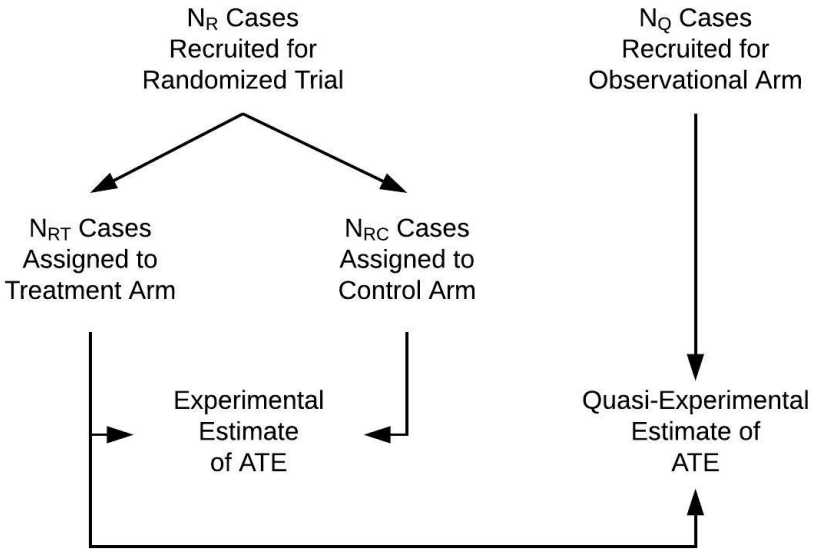

Figure 1. Figure illustrating the group structure of the three-arm design replication study. ATE = average treatment effect.

The four-arm design handles, through randomization, potential confounds that could invalidate comparisons based on three-arm designs; however, the four-arm design does not

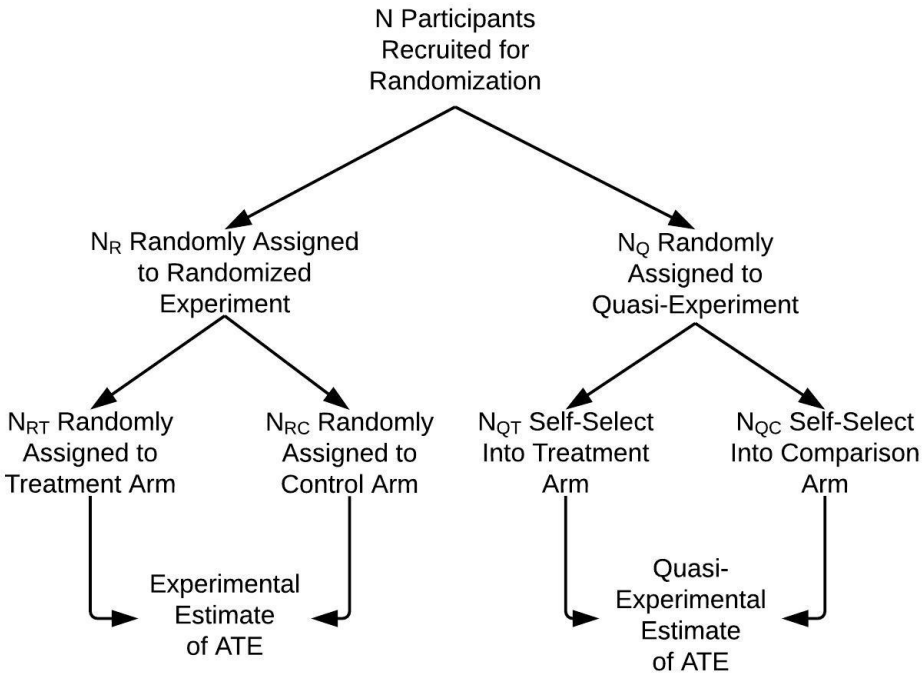

Figure 2. Figure illustrating the four-arm design of Shadish, Clark, & Steiner's (2008) design replication study. ATE = average treatment effect.

allow for estimation of conditional average treatment effects (ATEs) such as the average treatment effect on the treated (ATT), which are often of more interest in practice than the overall ATE (Little, Long, & Lin, 2008; Hill, 2008).

Purpose/Objective/Research Question

In this paper we report on the design, analysis, and results of a six-arm design replication study that incorporates an additional level of random assignment, suggested by Hill (2008) and developed and operationalized by Shadish and Steiner (2008), that permits experimental and quasi-experimental estimation of the ATE, ATT, and average treatment effect on the controls (ATC); see Figure 3.

The experimental arm of the six arm design is identical to the experimental arms of the three- and four-arm design replication studies. On the quasi-experimental side,

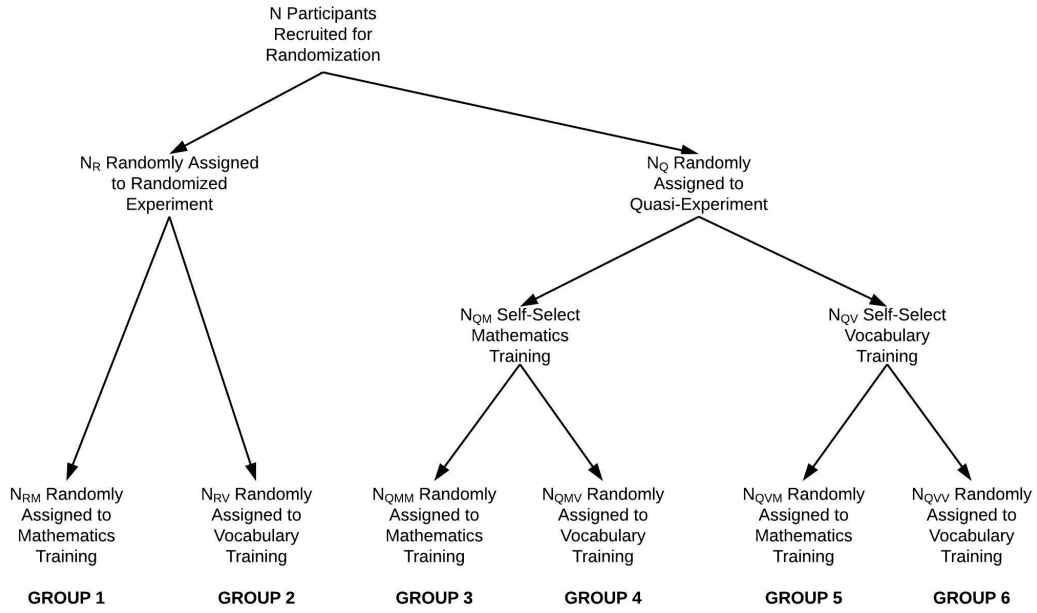

Figure 3. Figure illustrating the six-arm design proposed by Shadish & Steiner (2008)

however, participants are asked to select a training, either mathematics or vocabulary, as in Shadish et al. (2008). Then, no matter which training was selected, an additional randomization determines the actual training received. The purpose of asking participants to select training but then randomizing assignment despite their selection is to sort them into groups based on which training they would have selected. This is the key innovation that enables estimation of conditional average treatment effects for the treated and untreated groups.

Population/Participants/Subjects

Study participants are recruited online through Amazon's Mechanical Turk (MTurk) marketplace. MTurk connects workers with employers in need of completed tasks that require human intelligence. The MTurk marketplace was created in 2005 and has since grown to include hundreds of thousands of workers, primarily from the U.S. and India (Difallah, Filatova, & Ipeirotis, 2018). The workers are 65% female, have a modal annual household income between $40,000 and $60,000, 60% are older than 30, and over 75% hold

a bachelor's degree or higher (Barger, Behrend, Sharek, & Sinar, 2011). Workers can be sent an Mturk task that links to surveys created with software like Qualtrics and it is possible to receive thousands of completed surveys in a matter of hours. The quality of survey data obtained through MTurk is typically as reliable than those obtained by more traditional methods (Buhrmester, Kwang, & Gosling, 2011).

The benefit to using an online platform such as MTurk for participant recruitment is the potential to quickly enroll and gather data on a vast numbers of participants. Because our interventions target mathematics and vocabulary and assume a baseline level of competence, we require that all our participants are U.S. high school graduates.

Intervention/Program/Practice

In line with Shadish et al. (2008), we collect baseline measures of demographic information, math and vocabulary aptitude, social and emotional health, and preferences related to mathematics and vocabulary. The study design results in the creation of six groups, as detailed in Figure 3. Participants assigned to groups 1, 3, or 5 receive the mathematics training, whereas participants assigned to groups 2, 4, or 6 receive the vocabulary training.

The mathematics intervention consists of a short training on laws of exponents, including examples demonstrating product and quotient rules and exponentiation rules. The vocabulary intervention consists of a short training wherein participants study a list of vocabulary words. Directly after the training, all participants take both the mathematics and the vocabulary posttests. Both posttests feature some items that were present in the training materials and some that were not. We expect the treatment effect to be driven primarily by the shared items also present in the the training, and selection bias in the quasi-experiments to be driven primarily by the unique items not present in the training.

Data Collection and Analysis

The six arm design allows for three estimands to be targeted based on randomized and quasi-experimental designs: the ATE, ATT and ATC,

$$\begin{aligned}
\text{ATE} &= E[Y_i^1 - Y_i^0], \\
\text{ATT} &= E[Y_i^1 - Y_i^0 | T_i = 1], \\
\text{ATC} &= E[Y_i^1 - Y_i^0 | T_i = 0],
\end{aligned}$$

where Y_i^1 and Y_i^0 are the potential outcomes for unit i under treatment $T_i = 1$ and $T_i = 0$, respectively.

Consider first the effect of the mathematics intervention. Due to randomization, the differences in sample means between groups 1 and 2, 3 and 4, and 5 and 6, respectively represent unbiased estimators for the ATE, ATT, and ATC. For quasi-experiments, group 3 represents those who self-selected mathematics training and received mathematics training, and group 6 represents those who self-selected vocabulary training and received vocabulary training. Thus, a quasi-experimental comparison for each estimator may be constructed based on groups 3 and 6. The vocabulary intervention is handled similarly.

Estimates based on the randomized experiments are constructed both with and without additional covariance adjustment. For the quasi-experimental estimates, a number of causal estimators are explored including multiple regression analysis, propensity score analysis (Rosenbaum & Rubin, 1983), regression estimation (Schafer & Kang, 2008), and G-computation (Austin, 2012).

Findings/Results

Estimates from the randomized benchmarks and corresponding quasi-experiments will be presented and compared. Covariate balance, propensity score overlap, and graphical diagnostics will be presented for the quasi-experimental comparisons. From a replication

standpoint, we will be interested in the pattern of discordance and concordance of estimates before and after implementation of conditioning strategies for the quasi-experiments. This will be measured as proportion of bias reduction and with statistical tests of equivalence, when appropriate.

Conclusions

To our best knowledge, this study represents the first design replication that allows for comparisons of experimental and quasi-experimental estimates for conditional ATEs such as the ATT, which is typically of greater interest to educational researchers than the overall ATE. Practical considerations for the implementation of six-arm design replications studies, and trade-offs with respect to estimation of conditional ATEs will be discussed.

References

Austin, P. C. (2012). Using ensemble-based methods for directly estimating causal effects: An investigation of tree-based g-computation. *Multivariate Behavioral Research*, *47*, 115–135.

Barger, P., Behrend, T. S., Sharek, D. J., & Sinar, E. F. (2011). I-O and the crowd: Frequently asked questions about using Mechanical Turk for research. *The Industrial-Organizational Psychologist*, *49*, 11.

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's mechanical turk: A new source of inexpesnive, yet high-quality, data? *Perspectives on Psychological Science*, *6*, 3–5.

Cook, T. D., Shadish, W. R., & Wong, V. C. (2008). Three conditions under which observational studies produce comparable causal estimates: New findings from within–study comparisons. *Journal of Policy Analysis and Management*, *27*, 724–750.

Difallah, D., Filatova, E., & Ipeirotis, P. (2018). Demographics and dynamics of mechanical turk workers. *Proceedings of WSDM 2018: The Eleventh ACM International Conference on Web Search and Data Mining , Marina Del Rey, CA, USA, February 5âĂŞ9, 2018*, *103*, 9 pages.

Hill, J. (2008). Comment on can nonrandomized experiments yield accurate answers? a randomized experiment comparing random and nonrandom assignments. *Journal of the American Statistical Association*, *103*, 1346-1350.

Little, R. J., Long, Q., & Lin, X. (2008). Comment on can nonrandomized experiments yield accurate answers? a randomized experiment comparing random and nonrandom assignments. *Journal of the American Statistical Association*, *103*, 1344-1346.

Pohl, S., Steiner, P. M., Eisermann, J., Soellner, R., & Cook, T. D. (2009). Unbiased causal inference from an observational study: Results of a within-study comparison. *Educational Evaluation and Policy Analysis*, *31*, 463–479.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*, 41-55.

Schafer, J., & Kang, J. (2008). Average causal effects from nonrandomized studies: A practical guide and simulated example. *Psychological Methods*, *13*, 279–313.

Shadish, W. R., Clark, M. H., & Steiner, P. (2008). Can nonrandomized experiments yield accurate answers? a randomized experiment comparing random and nonrandom assignments. *Journal of the American Statistical Association*, *103*, 1334–1343.

Shadish, W. R., & Steiner, P. M. (2008). *A laboratory analogue method for validating statistical adjustments to nonrandomized experiments.* (Unpublished Grant Proposal)