Symposium Information

Organizer: Vivian C. Wong (University of Virginia)
Contact email: vcw2n@virginia.edu
Title: Towards a Reproducibility Framework
First choice of conference session: Research Methods

Paper 1

Title: Improving the Science in Replication Sciences
Authors: Vivian C. Wong (University of Virginia) and Peter M. Steiner (University of Wisconsin Madison)

Vivian Wong*: vcw2n@virginia.edu
Peter M. Steiner: psteiner@wisc.edu

Paper 2

Title: Replication and Robustness in Educational Research
Authors: Mimi Engel, University of Colorado Boulder; Amy Claessens, University of Chicago; Sarah Kabourek, Vanderbilt University

Mimi Engel*: mimi.engel@colorado.edu
Amy Claessens: amyc1@uchicago.edu
Sarah Kabourek: sarah.e.kabourek@vanderbilt.edu

Paper 3

Title: Open Science for Education Science: Toward Transparent, Reproducible Workflows for Intervention Research
Author: Sean Grant (RAND Corporation)

Sean Grant*: sgrant@rand.org

Discussant

Jessaca Spybrook (Western Michigan University)

Jessaca Spybrook*: jessaca.spybrook@wmich.edu

* Corresponding authors indicated in asterisks.

Symposium Justification
Title: Towards a Reproducibility Framework

Background

Recent efforts to promote evidence-based practices in medicine and the social sciences (i.e. What Works Clearinghouse) assume that scientific findings are of sufficient validity to warrant its use in decision making. Reproducibility 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 is established through systematic, transparent, and reproducible methods, results that are independently verified and replicated, and findings that are generalizable to at least some target population of interest.

Given the central role of replication and reproducibility in the accumulation of scientific knowledge, recent methodological work has examined both the prevalence and success of replicating seemingly well-established findings. Thus far, results from these replication efforts have not been promising. The Open Science Collaboration (OSC) conducted replications of 100 experimental and correlational studies published in high impact psychology journals. Overall, the OSC found that only 36% of the replication studies produced results with the same statistical significance pattern as the original study. These findings prompted the OSC authors to conclude that replicability rates in psychology were low, but not inconsistent with what has been found in other domains of science. In 2005, Ioannidis argued that most findings published in the biomedical sciences were likely false (2005). His review of more than 1,000 medical publications found that only 44% of replication studies produced results that corresponded with the original findings (2008). Combined, these results contribute to a growing sense of a "replication crisis" occurring in multiple domains of science.

Proposed Symposium

Considerable disagreement remains about what replication is, its role in science, and what – if any – institutional supports are needed to promote replication efforts. This symposium addresses these issues. ***The first paper presents a formal definition of the replication design*** using a potential outcomes framework. It describes five stringent assumptions required for two studies to yield identical results (within the limits of sampling error), and demonstrates the advantages and limitations of different replication design variants for addressing these assumptions. ***The second paper examines the prevalence of replication efforts in education settings***. The authors look at the percentage of studies that examine the robustness of treatment effects across different datasets, methods, and subgroups in two high impact education journals over two years (1994, and 2014). ***The third paper provides guidance for promoting a transparent, reproducible workflow in education evaluation research.*** The framework recommends: 1) pre-registration in the design phase, 2) a data management plan in the analysis phase, 3) a clear reporting protocol in the dissemination phase, and 4) a system for archiving data and analysis code. ***The symposium will conclude with comments from a discussant, who will address SREE's recent efforts to promote data and methods transparency, as well as comment on papers presented in the session.***

References

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science, 349*(6251), aac4716.

Ioannidis JPA (2005) Why Most Published Research Findings Are False. PLoS Med 2(8): e124. https://doi.org/10.1371/journal.pmed.0020124

Title: Improving the Science in Replication Sciences

Background

Considerable attention has been devoted to examining the prevalence and success of reproducibility efforts. Despite consensus to promote the reproducibility of scientific findings, there is substantial disagreement about how these results should be interpreted. There are three reasons why study results may not reproduce. They include random error, low statistical power, and bias (Gilbert, King, Pettigrew, et al., 2016). Recent methodological work on reproducibility has focused on statistical issues related to error and power for detecting comparability of results (Benjamin, Berger, Johannesson et al., 2017; Simonsohn, 2015). Bias, however, remains a serious challenge.

Bias refers to differences in the original and replication studies that is related to the study's outcome. In replication contexts, potential sources of bias are numerous and broadly defined. This includes when researchers manipulate or selectively present findings, when there are deviations in the research protocol, when treatment and control conditions vary, and/or when the method of analysis changes across studies. Contextual changes that occur between the original and replication studies may also affect the reproducibility of results. Combined, these criticisms suggest that bias may be the key methodological challenge for replication, especially in fields where there is high uncertainty in the outcome and limited experimenter control.

This paper provides a formal understanding of replication as a research design. We describe five assumptions needed for two studies to produce identical results (within the limits of sampling error). As we will see, assumptions for replication in field settings are stringent and often are not feasible. However, by employing clever research design elements and empirical diagnostics for ruling out plausible biases, replication approaches may be a useful tool for evaluating the reproducibility of methods and scientific claims, and for examining the robustness of treatment over systematic sources of variation. To this end, this paper highlights two design variants of replication, and describes their strengths and weaknesses.

Conceptual Framework

Figure 1 describes a replication design with two study arms – an "original" and a "replication" study. The researcher compares findings from one study to results obtained from the second study to assess the reproducibility of results. Here, participants are sampled from an overall target population into study arm 1 (i.e. the original study) or study arm 2 (i.e. the replication study). Within each study arm, participants are assigned again – or select into – a treatment or control condition. If results from both study arms are judged to be sufficiently "close" (i.e. in terms of direction, size, and/or statistical significance patterns), the researcher concludes that the finding has been reproduced.

Five assumptions are required for direct replication of results.[1] The first is a **stable-unit-treatment-value assumption (SUTVA)** (A1). This requires that the treatment and control conditions are well defined, and that they are equivalent across study arms. The assumption is violated if there are variations in treatment and control conditions across the two study arms, such as when treatment protocols differ, when the original study has a stronger treatment dosage than the replication, or when outcome measures differ across study conditions. The second assumption requires participants' potential outcomes are **independent or conditionally independent of their observed assignment into study arms** (A2). The assumption implies that there are no differences in participants across the two study arms that also are related to their potential outcomes (i.e. participants in the original study are not more advantaged than those in the replication arm). The third (A3) and fourth (A4) assumptions require that within each study arm, **participants' potential outcomes are independent of their treatment status**. In program evaluation, this assumption is often met through a randomization procedure into treatment conditions, such as a fair coin toss. The fifth assumption states that treatment effects are estimated using an **unbiased analytic procedure**, such as regression adjustment, and that requirements for the estimation procedure are met (i.e. no misspecification of functional form) (A5). When these five design assumptions are met, treatment effects should be identical across both study arms (within the limits of sampling error). Different treatment effects imply that one or more design assumption has been violated, or researcher error in reporting or analysis.

[1] See Methodological Appendix A for formal definition and proof.

Replication Design Variants

Although assumptions for replication are stringent, the researcher may adopt various design approaches to evaluate the reproducibility of results. In **prospective designs**, both study arms – the original and replication – are conducted simultaneously. This allows the researcher to incorporate design features as well as empirical diagnostic measures to meet replication assumptions. To address A2, participants may be randomly assigned into the two study arms, and within each arm, they may be assigned again into treatment and control conditions (A4 and A5). Because the replication is planned prospectively, the researcher may take steps to ensure that the same treatment protocol is implemented in both studies, that the same fidelity measures are used to monitor conditions, and the same outcomes are assessed. Contextual changes across study arms should not be a threat because treatments are implemented at the same time, and participants are drawn from the same target population. This helps address the SUTVA requirement. Finally, treatment effects from both studies may be analyzed simultaneously, using the same unbiased analytic procedures. A variant of this approach was introduced by Shadish et al. (2008), when university students were randomly assigned into two study arms, and assigned again into a short vocabulary or math intervention. The interventions were implemented in the same way in the first and second study, and treatment effects were compared across the two study arms to determine it's reproducibility.

In **matched designs**, the two studies are conducted at different time periods, often by independent investigators. The goal here is for an independent investigator to "match" conditions from the original study, taking care to meet all five replication assumptions. This includes using the same eligibility criteria and sampling procedures to recruit participants from the same target population (A2); implementing the same treatment and control protocols in similar settings, collecting information on the same outcome measures (A1); and using the same research design and analysis strategy for producing results (A3, A4, and A5). However, results from matched designs can be challenging to interpret because bias may occur if *any* of the design assumptions are violated. Specifically, if there are deviations in the treatment protocol, differences in the sample and context that change over time, or investigator error. The Open Science Collaborative adopted this approach in their effort to replicate 100 well-known findings in psychology. They judged that only 39% of results replicated (Novack et al., 2015).

Implications

Direct replications have stringent assumptions for producing interpretable results. However, there are well-established methods for addressing the assumptions described above. The most compelling approaches will involve a combination of research designs for ruling out bias threats, and diagnostics for probing assumptions empirically.

References

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E.J., Berk, R., ... Johnson, V. E. (2017). Redefine statistical significance. *Nature Human Behaviour*.

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

Gilbert, D. T., King, G., Pettigrew, S., and Wilson, T. D. (2016). Comment on "Estimating the reproducibility of psychological science." *Science*, 351(6277), 1037 LP-1037.

Shadish, W. R., Clark, M. H., and Steiner, P. M. (2008). Can Nonrandomized Experiments Yield Accurate Answers? A Randomized Experiment Comparing Random and Nonrandom Assignments. *Journal of the American Statistical Association*, 103(484), 1334–1344.

Simonsohn, U. (2015). Small Telescopes. *Psychological Science*, 26(5), 559–569.

Figures

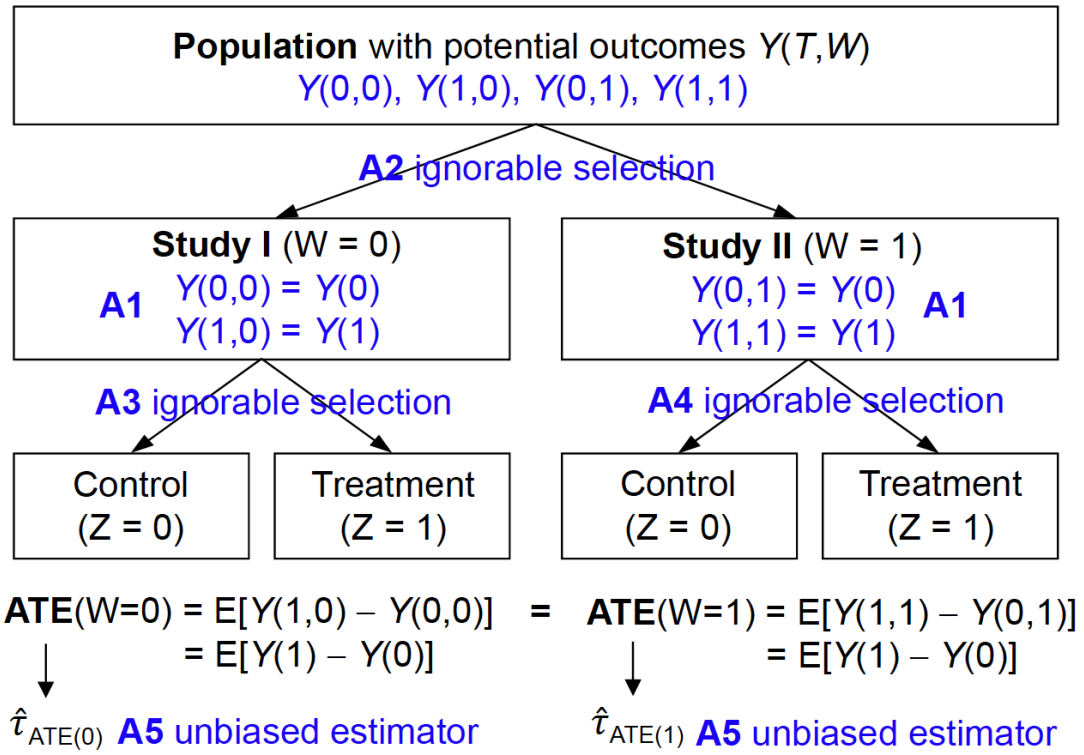


Figure 1: Replication Design and Required Assumptions

Methodological Appendix A

We begin by considering two studies, the original and replicated study $W_i \in \{0, 1\}$, where W_i is coded 0 if participant i belongs to the original study arm, and 1 if the participant is in the replicated study condition. Within each study arm, there are two treatment conditions $T_i \in \{0, 1\}$ if the participant is in the control condition, and 1 if it is in the treatment. Under this conceptualization, each participant i has four potential outcomes, $Y_i(T_i = t, W_i = w)$, which include control and treatment outcomes in the original study, $Y_i(0, 0)$ and $Y_i(1, 0)$, and control and treatment outcomes in the replication study, $Y_i(0, 1)$ and $Y_i(1, 1)$.
With this framework in mind, we define the causal estimands of interest for the original and replicated study. For clarity in notation, we assume that both the original and replicated studies to be randomized controlled trials (RCTs). For the original study ($W = 0$), the causal estimand of interest is the average treatment effect $ATE(W = 0)$ for the target population:

$$ATE(0) = E(Y_i(1, 0)) - E(Y_i(0, 0)) = E(Y_i(1, 0) - Y_i(0, 0)).$$

Here, the expectation is taken over all participants in the original and replication study arms, or the Average Treatment Effect (ATE) if all participants are assigned to the original study. Correspondingly, we define the $ATE(1)$ for the replication study ($W = 1$) as:

$$ATE(1) = E(Y_i(1, 1)) - E(Y_i(0, 1)) = E(Y_i(1, 1) - Y_i(0, 1)),$$

where again, the expectation is taken over the entire population.

Assumptions for a Direct Replication Design

Five assumptions are required for the original and replication study to yield comparable results. The first assumption is a Stable-Unit-Value-Assumption, which ensures that both study designs produce equivalent causal estimands of interest. Assumptions 2, 3, and 4 imply ignorable assignment across study arms (original and replication studies), and into treatment conditions within each study arm. Finally, the fifth assumption states that treatment effects are estimated using unbiased analytic procedures.[2] When these five assumptions are met, then the original and replication study produces the same treatment effect. Any observed differences in effect estimates across study arms is the result of sampling error.

A1. Stable-Unit-Value-Assumption (SUTVA)

Although these SUTVA is standard requirement in experimental, quasi- and non-experimental research designs, this assumption has special applications in the replication context. Specifically, SUTVA implies:

1. *Non-interference.* Participants' potential outcomes depend only on its own study arm and treatment status, and not the status of others:

   $$Y_i(T_i, T_{-i}, W_i, W_{-i}) = Y_i(T_i, W_i)$$

   where T_{-i} and W_{-i} are the assignment vectors of all other study participants except for participant i.

2. *Excludability.* The study condition and treatment statuses, W and T, are uniquely defined, such that there are no different versions of each condition (Rubin, 1980; Rubin, 1986). Moreover, the potential control and treatment outcomes do not depend on study condition status. Formally,

   $$Y_i(T_i, W_i = w, Z_i = z, G_i = g) = Y_i(T_i, W_i = w', Z_i = z', G_i = g') = Y_i(T_i)$$

   with $w \neq w'$, $z \neq z'$ and $g \neq g'$. Here, G is a third variable with respect to study arm conditions, and Z a corresponding third variable with respect to the treatment conditions within each study arm (original or replication study).

[2] A weaker version of this assumption requires only that the same analytic procedure is used across both study arms.

The first implication of SUTVA states that participants' potential outcomes are unaffected by others' assignment to the original or replication study conditions, nor do they react to the treatment status of others within each study arm. In the replication context, this requirement may be weakened to allow for interference between participants, but not differential interference across units.
The second implication of SUTVA that there are no hidden variations in treatment and control status within and across study condition arms, and that assignment to the original and replication study has no effect on participants' potential outcomes. In replication contexts, this requirement has special implications for researchers to consider. For example, one way in which the assumption may be violated is if research protocol varies across the original and replication study designs. Another possible violation occurs if the original study has a stronger treatment dosage or if comparisons in the replication study have alternative treatment options not available to control units in the original study. Finally, if participants are aware of results from the original study, they may respond differentially to treatment and control conditions.

Causal Estimand of Interest in Replication Designs with RCTs

When requirements for SUTVA (non-interference and excludability) are met, then the causal estimand in the original and replication study is identical, such that $ATE(0) = ATE(1)$, and the observed outcome is a function of each participant's potential outcomes, and its status T. Thus, for each participant i, the observed outcome may be written as:

$$Y_i = Y_i(0,0)(1-T_i)(1-W_i) + Y_i(0,1)(1-T_i)W_i + Y_i(1,0)T_i(1-W_i) + Y_i(1,1)T_iW_i,$$

where the second line follows from the exclusion restriction, which implies that
$Y_i(0,1) = Y_i(0,0) = Y_i(0)$ and $Y_i(1,1) = Y_i(1,0) = Y_i(1)$.
Thus far, we have defined $ATE(0)$ and $ATE(1)$ in terms of their expected potential outcomes. In practice, these quantities cannot be computed because the researcher observes only one potential outcome for each participant. With three additional assumptions, we can show that the conditional expectations of the four observed outcomes, $E(Y_i|T_i = 0, W_i = 0)$, $E(Y_i|T_i = 1, W_i = 0)$, $E(Y_i|T_i = 0, W_i = 1)$, and $E(Y_i|T_i = 1, W_i = 1)$, may be used to identify ATEs for the original and replication study conditions. Because potential outcomes depend only on T (by excludability), we no longer index them by the study condition status, W. Below, we discuss three assumptions required for identification of treatment effects across both and within each study arms.

A2. Ignorable Selection into the Original and Replication Study Designs

This assumption requires that participants' potential outcomes are independent or conditionally independent of their observed assignment to the original and replication study conditions, such that: $(Y_i(0), Y_i(1))\perp W_i$.
This assumption implies that knowledge of participants' potential outcomes provide no information about their status in the original or replication study design. In practice, this requirement is rarely met in field settings given that it requires random assignment of units into the original and replicated study. Generalization work from Tipton et al. and Stuart et al. is based on this assumption. Their efforts are to match comparable participants across the original and replication studies.

A3. Ignorable Selection into Treatment Conditions in the Original Study

For participants in the original study (W = 0), potential outcomes are independent of treatment, such that: $(Y_i(0), Y_i(1))\perp(T_i|W_i = 0)$. In an RCT, this assumption is through random assignment of participants into treatment conditions of the original study. In QE or NE studies, other design assumptions for identification may be needed, such as the common trend assumption in DID, continuity assumption in potential outcomes (or, local randomization) in RD, or conditional independence assumption in an observational study.

A4. Ignorable Selection into Treatment Conditions in the Replication Study

For participants in the replicated study (W = 1), potential outcomes must also be independent of treatment such that: $(Y_i(0), Y_i(1))\perp(T_i|W_i = 1)$. Similar to assumption 3 from above, this requirement may be met through random assignment in the replication study design, or through other identification assumptions in quasi- or non-experimental designs.

A5. Unbiased Estimator of Treatment Effects Across Both Study Arms

Even when the above identification assumptions are met, for an original and replication study to produce reproducible treatment effects requires that results are estimated through an unbiased analytic procedure, or identical analytic procedures (that may be biased, but not differentially biased across study conditions). For example, if parametric regression is used to estimate treatment effects, the functional form of the independent and dependent variable must be correctly specified. Generally, treatment effects should be estimated using the same analytic procedures, ensuring that even if the estimator is biased, the bias should not occur differentially across study conditions.

Nonparametric Identification of the ATE in Replication Designs

Given assumptions A1-A5, we first show that $ATE(0)$ and $ATE(1)$ are identified and then show that they are equivalent. Under assumptions A1-A4, the difference in the expected observed outcomes, $\tau_1(0)$, can be used to identify ATE(0) in the RCT benchmark:

$$\begin{aligned}
\tau_1(0) &= E(Y_i|T_i = 1, W_i = 0) - E(Y_i|T_i = 0, W_i = 0) \\
&= E(Y_i(1)|T_i = 1, W_i = 0) - E(Y_i(0)|T_i = 0, W_i = 0) \\
&= E(Y_i(1)|W_i = 0) - E(Y_i(0)|W_i = 0) \\
&= E(Y_i(1)) - E(Y_i(0)) = ATE(0)
\end{aligned}$$

Line 1 is the difference in the expectations of the observed treatment and control outcomes in the RCT benchmark. Line 2 follows from SUTVA (A1 and A2), line 3 from the independence between potential treatment and control outcomes (A4), and line 4 from the independence of WSC status and potential outcomes (A3).
If A1-A3 and A5 are met, the difference in the expected observed outcomes, $\tau_1(1)$, identifies $ATE(1)$ in the NE arm (non-equivalent comparison group design):

$$\begin{aligned}
\tau_1(1) &= E_X\big\{E(Y_i|T_i = 1, W_i = 1, X_i = x) - E_X(Y_i|T_i = 0, W_i = 1, X_i = x)\big\} \\
&= E_X\big\{E(Y_i(1)|T_i = 1, W_i = 1, X_i = x) - E_X(Y_i(0)|T_i = 0, W_i = 1, X_i = x)\big\} \\
&= E_X\big\{E(Y_i(1)|W_i = 1, X_i = x) - E_X(Y_i(0)|W_i = 1, X_i = x)\big\} \\
&= E(Y_i(1)|W_i = 1) - E(Y_i(0)|W_i = 1) \\
&= E(Y_i(1)) - E(Y_i(0)) = ATE(1)
\end{aligned}$$

Line 1 is the difference in the expectations of the observed outcomes for treatment and comparison cases in the non-experimental arm. Line 2 is equivalent due to SUTVA (A1 and A2), line 3 follows from A5 (strong ignorability in the replication), line 4 takes the expectation over the distribution of X, and line 5 follows from A3 (independence of WSC conditions).
Note that the expectations in the last lines (unconditional on W) are taken over units in the original and replication arms together. This produces the ATE of the entire inference population. In addition, because assumption A2 implies $Y_i(0,1) = Y_i(0,0) = Y_i(0)$ and $Y_i(1,1) = Y_i(1,0) = Y_i(1)$ (used in lines 2), the last lines of both proofs are identical, such that: $ATE(0) = ATE(1)$.

Title: Replication and Robustness in Educational Research

Background

Whether results from a single study are replicable across contexts and over time is a key question related to the scientific method. Further, robustness checks – examining the extent to which findings from a single study are consistent across subgroups and analytic approaches – are an important means for determining whether a quantitative result 'holds up' under careful scrutiny or is fragile and, therefore, more likely to be spurious, are more common in some academic disciplines than others. Recent research finds that published studies that emphasize replication are rare in both applied economics and developmental psychology. Researchers analyzed published articles in top field journals in the two disciplines, comparing recent publications with articles published two decades prior. In addition to finding that articles replicating prior research were extremely uncommon in both applied economics and developmental psychology, that study found robustness checking techniques to be much more commonly used in applied economics than in developmental psychology (Duncan, et al., 2014). The purpose of the current study is to examine the extent to which replication and robustness practices are used in empirical educational research studies published in peer reviewed journals.

Research Questions

The current study expands upon this prior research to examine how common replication practices and robustness checks are in empirical studies published in peer reviewed education research journals. We apply the coding scheme developed by Duncan and colleagues (2014) to answer the following research questions:

1) How often do articles published in high impact educational research journals in 2014 and, for comparison, in 1994, include replication of prior research?
2) How often do articles published high impact educational research journals in 2014 and, for comparison, in 1994, include robustness checks such as the use of multiple data sets, subgroup analyses, and multiple estimation techniques?

Research Methods

The current study is an empirical investigation examining the use of replication and robustness methods in educational research. We compare results from our analyses with results from recent research in applied economics and developmental psychology. Two top field journals in education, both journals of the American Educational Research Association, were coded; *American Educational Research Journal (AERJ)*, and *Educational Evaluation and Policy Analysis (EEPA)*. *AERJ* was selected because it is considered a flagship journal for the broad field of education and evidence of replication and robustness in that journal provides information about the extent to which these practices are valued in the field as a whole. *EEPA* was selected because it is a top field journal for applied quantitative analysis in education and because we anticipated that it would represent a case where robustness checking practices would more likely be encouraged.

Following the methods used in related research (Duncan, et al., 2014), we coded a total of 200 articles, half of which were published in 1994 and half in 2014. Articles were coded by the authors and graduate students with knowledge of quantitative research methods and who were trained by the authors in the explicit coding methodology. Articles that were theoretical or purely qualitative studies (e.g., ethnographies) were excluded from the sample. Each eligible article was coded to examine 1) whether it included efforts to replicate results from prior research, and 2) whether it included robustness checking techniques including a) the use of two or more data sets, b) the use of multiple analytic or estimation techniques, and c) subgroup analyses aimed at determining whether main results held for subpopulations (e.g., gender, race/ethnicity).

Results

Table 1 presents a brief description of the articles included in the analyses. As shown, nearly one quarter of articles published in both journals in 2014 utilized public use data sets—creating the possibility for others to attempt to replicate the published findings. A similar fraction of articles in *EEPA* also used public use data in 1994, while only 6% of articles in *AERJ* from 1994 used publicly available data. The fraction of articles using a random assignment design is similar in both journals in 2014, while fewer, only 2% of articles, studies using random assignment were published in *EEPA* in 1994.

Table 2 presents the preliminary results for the replication and robustness checks in both journals at each time point. As shown, in the articles examined, explicit replication of prior work almost never occurs—just 2% in *AERJ* 1994 and 2% in *EEPA* 2014. In terms of robustness checks, few of the articles in *AERJ* in 1994 or 2014 use multiple datasets or estimation techniques. This is also true in *EEPA* in 1994. In terms of subgroup replication, the fraction of articles that do this is higher in *AERJ* 2014 than 1994 (14% versus 2%). Articles in *EEPA* in 2014 are the most likely to include at least one of the robustness checking techniques. The last column of Table 2 shows the fraction of articles in which authors conducted any of the robustness checks. Few of the empirical articles coded included any robustness checking in 1994, while nearly half of the *EEPA* articles from 2014 included at least one of these techniques. Preliminary results based on evidence of what was published across 200 articles suggests that replication of prior research is not highly valued in educational research. Robustness checks are somewhat more common, particularly in EEPA, but are not ubiquitous in these two top field journals.

Implications

Based on its extremely low incidence across 550 articles in six peer-reviewed journals (Duncan et al., 2014 and the results presented here), we see little evidence suggesting that replication is encouraged in top journals in three fields—education, economics, and developmental psychology. We recommend that education research journal editorial boards adopt guidelines that encourage both replication and robustness checking techniques, as is being done in other fields and disciplines.

Reference

Duncan, G. J., Engel, M., Claessens, A., & Dowsett, C. J. (2014). Replication and robustness in developmental research. *Developmental psychology*, *50*(11), 2417.

Table 1: Descriptive characteristics of coded articles

Journal	Period	Number of articles coded	Number of non-empirical articles not coded	Public use datasets	Random assignment to treatment/control conditions
American Educational Research Journal	Current	50	19	24%	14%
	20 years ago	50	50	6%	18%
Educational Evaluation and Policy Analysis	Current	50	3	24%	20%
	20 years ago	50	33	22%	2%

Table 2: Replication and robustness-checking practices

Journal	Period	Meta-Analysis	Explicit replication of prior research	Robustness checking practices			Any replication or robustness-checks
				Two or more data sets	Two or more estimation techniques	Subgroup replication	
American Educational Research Journal	Current	0%	0%	2%	0%	14%	16%
	20 years ago	0%	2%	4%	0%	2%	8%
Educational Evaluation and Policy Analysis	Current	2%	2%	16%	10%	36%	50%
	20 years ago	2%	0%	4%	0%	4%	8%

Title: Open Science for Education Science:
Toward Transparent, Reproducible Workflows for Intervention Research

Background

The scientific community has created a reward system that does not sufficiently incentivize the vital features of science: i.e., transparency, openness, and reproducibility (McNutt, 2014). Concerns about research waste, scientific misconduct, and lack of replication are consequently rising (Glasziou et al., 2014; Open Science Collaboration, 2012, 2015; Tajika, Ogawa, Takeshima, Hayasaka, & Furukawa, 2015). For instance, a project to replicate 100 experimental and correlational studies in psychology found that only 36% of replications (versus 97% of original studies) had significant results, and the mean effect size of replications was half the magnitude of the mean effect size of the original effects (Open Science Collaboration, 2012, 2015). A similar attempt to replicate 67 papers in economics was only able to replicate less than half of the papers, even with help from the original authors (Chang & Li, 2015). To improve the credibility of the scientific enterprise, researchers have begun to list, develop, and implement "open practices" that can help increase the transparency and reproducibility of research workflows.

Objective

This presentation will introduce current best practices in open science to education scientists who conduct research on the effects of educational interventions. This presentation has two overall goals: (1) to convince attendees to incorporate best practices of open science in future projects that evaluate educational interventions, and (2) to motivate attendees to promote the use of best practices in open science to other stakeholders of educational intervention research.

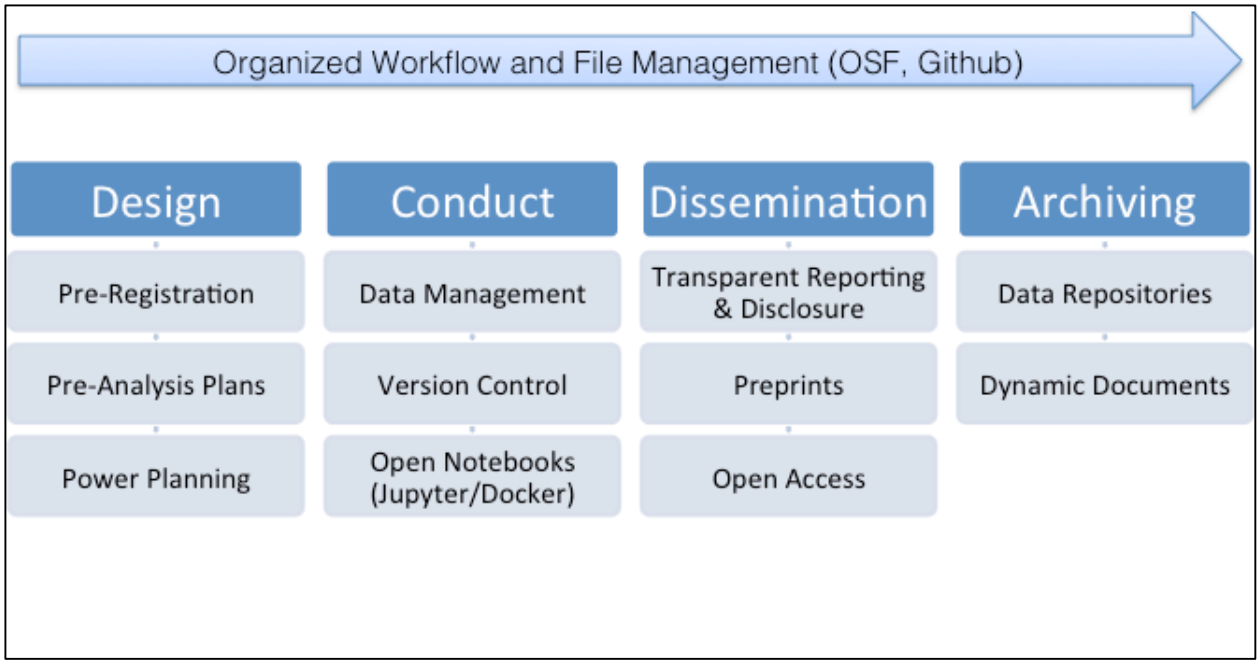
Research Design

To achieve the above objectives, this presentation will discuss a conceptual framework for transparent, reproducible workflows in educational intervention research, based on the work of several groups that focus on research transparency, namely: the *Lancet* REWARD Campaign (Moher et al., 2015); the Center for Open Science and their Transparency and Openness Promotion (TOP) Guidelines (Nosek et al., 2015); the Berkeley Initiative for Transparency in the Social Sciences (Miguel et al., 2014); the Data Access and Research Transparency (DA-RT) group (Data Access and Research Transparency group, 2015); the Meta-Research Innovation Center at Stanford (Ioannidis, Fanelli, Dunne, & Goodman, 2015); and the Laura and John Arnold Foundation (Preston, 2011).

Findings

A transparent, reproducible workflow for education intervention research involves an organized set of practices during study design, conduct, dissemination, and archiving (Figure 1).

Figure 1. A Transparent, Reproducible Workflow for Educational Intervention Research



As part of design, education researchers should *pre-register their studies*, specifying in as much details as possible the plans for each study — such as recruitment strategies, eligibility criteria, intervention procedures, measurement plans, and operationalized hypotheses — in public, time-stamped, and locked online repositories (Zarin, Tse, Williams, Califf, & Ide, 2011). Particularly for studies evaluating the effects of interventions (e.g., randomized trials), pre-registrations should include *pre-analysis plans* (Olken, 2015) — step-by-step plans specifically setting out how the researchers will analyze the data. These documents should include details on *plans to ensure adequate statistical power* for detecting minimally important effect sizes (Benjamin et al., 2017). As part of conduct, education researchers should ensure *proper management of their data* using bespoke tools such as the Open Science Framework (Foster & Deardorff, 2017) that incorporate *version control software* (recording changes to files over time to recall specific versions later) and *open notebooks* (public sharing of the primary records of research studies). As part of disseminating the results of studies, researchers should utilize reporting guidelines for *transparent reporting and disclosure* of study details, more immediately disseminate findings through *pre-print* servers and publications, and aim to publish their findings *open access* so that research outputs that are freely available to interested readers. As part of archiving studies, education researchers should share their data (with explanatory metadata) in trusted *data repositories* and record their analyses using *dynamic documents* (such as RMarkdown) that include all analytical code underpinning reported findings.

Conclusions

Researchers can use these open practices during each study evaluating educational interventions to ensure that their workflows are more transparent and reproducible. In addition, other stakeholders of education research can facilitate the adoption of open practices by educational intervention researchers (Shamseer et al., 2015). For example, journal editors can implement policies and procedures that require adherence to these practices for manuscripts that they publish. Peer-reviewers at these journals can help ensure that authors adhere to these journal

standards (Stevens et al., 2014). Research funders could also adopt open polices and procedures to increase the transparency and reproducibility of the research that they commission on educational interventions. Furthermore, policy-makers and practitioners can encourage the adoption of transparent and reproducible workflows to increase the chances that positive findings on educational interventions are actually replicable in "real-world" contexts. Lastly, faculty can incorporate the instruction on these practices into their courses and mentoring to build the capacity of the next generation of education researchers to produce transparent and reproducible research. Researchers, editors, peer-reviewers, funders, and educators can consult the resources page of the Berkeley Initiative for Transparency in the Social Sciences for up-to-date information on practicing open science (www.bitss.org/resources/).

References

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E. J., Berk, R., . . . Johnson, V. E. (2017). Redefine statistical significance. *Nature Human Behaviour*. doi: 10.1038/s41562-017-0189-z

Chang, A. C., & Li, P. (2015). *Is Economics Research Replicable? Sixty Published Papers from Thirteen Journals Say "Usually Not". Finance and Economics Discussion Series 2015-083*. Washington, DC: Board of Governors of the Federal Reserve System.

Data Access and Research Transparency group. (2015). Data Access and Research Transparency (DA-RT): A joint statement by political science journal editors. *Political Science Research and Methods, 3*, 421.

Foster, E. D., & Deardorff, A. (2017). Open Science Framework (OSF). *Journal of the Medical Library Association: JMLA, 105*(2), 203.

Glasziou, P., Altman, D. G., Bossuyt, P., Boutron, I., Clarke, M., Julious, S., . . . Wager, E. (2014). Reducing waste from incomplete or unusable reports of biomedical research. *The Lancet, 383*(9913), 267-276.

Ioannidis, J. P. A., Fanelli, D., Dunne, D. D., & Goodman, S. N. (2015). Meta-research: Evaluation and Improvement of Research Methods and Practices. . *PLoS Biology, 13*(10), e1002264.

McNutt, M. (2014). Reproducibility. *Science, 343*(6168), 229.

Miguel, E., Camerer, C., Casey, K., Cohen, J., Esterling, K. M., Gerber, A., & Van der Laan, M. (2014). Promoting transparency in social science research. *Science, 343*(6166), 30-31.

Moher, D., Glasziou, P., Chalmers, I., Nasser, M., Bossuyt, P. M. M., Korevaar, D. A., . . . Boutron, I. (2015). Increasing value and reducing waste in biomedical research: Who's listening? *The Lancet, Available online 27 September 2015*, doi: 10.1016/S0140-6736(1015)00307-00304.

Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., . . . Yarkoni, T. (2015). Promoting an open research culture. *Science, 348*, 1422-1425.

Olken, B. A. (2015). Promises and perils of pre-analysis plans. *The Journal of Economic Perspectives, 29*(3), 61-80.

Open Science Collaboration. (2012). An open, large-scale, collaborative effort to estimate the reproducibility of psychological science. *Perspectives on Psychological Science, 7*(6), 657-660.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science, 349*(6251), aac4716.

Preston, C. (2011). A Thirtysomething Billionaire Couple Take on Tough Issues Via Giving. *Chronicle of Philanthropy*.

Shamseer, L., Moher, D., Clarke, M., Ghersi, D., Liberati, A., Petticrew, M., & Stewart, L. A. (2015). Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015: Elaboration and explanation. *BMJ, 349*, g7647.

Stevens, A., Shamseer, L., Weinstein, E., Yazdi, F., Turner, L., Thielman, J., . . . Moher, D. (2014). Relation of completeness of reporting of health research to journals' endorsement of reporting guidelines: Systematic review. *BMJ, 348*, g3804.

Tajika, A., Ogawa, Y., Takeshima, N., Hayasaka, Y., & Furukawa, T. A. (2015). Replication and contradiction of highly cited research papers in psychiatry: 10-year follow-up. *British Journal of Psychiatry, Available online July 2015*, doi: 10.1192/bjp.bp.1113.143701

Zarin, D. A., Tse, T., Williams, R. J., Califf, R. M., & Ide, N. C. (2011). The ClinicalTrials. gov results database—update and key issues. *New England Journal of Medicine, 364*(9), 852-860.