Empirical Article

Advances in Methods and Practices in Psychological Science
2018, Vol. 1(1) 7–18

DOI: 10.1177/2515245917740427
www.psychologicalscience.org/AMPPS

Improving the Replicability of Psychological Science Through Pedagogy

Robert X. D. Hawkins, Eric N. Smith, Carolyn Au, Juan Miguel Arias, Rhia Catapano, Eric Hermann, Martin Keil, Andrew Lampinen, Sarah Raposo, Jesse Reynolds, Shima Salehi, Justin Salloum, Jed Tan, and Michael C. Frank
Department of Psychology, Stanford University

Abstract
Replications are important to science, but who will do them? One proposal is that students can conduct replications as part of their training. As a proof of concept for this idea, here we report a series of 11 preregistered replications of findings from the 2015 volume of *Psychological Science*, all conducted as part of a graduate-level course. As was expected given larger, more systematic prior efforts, the replications typically yielded effects that were smaller than the original ones: The modal outcome was partial support for the original claim. This work documents the challenges facing motivated students as they attempt to replicate previously published results on a first attempt. We describe the workflow and pedagogical methods that were used in the class and discuss implications both for the adoption of this pedagogical model and for replication research more broadly.

Keywords
replication, reproducibility, pedagogy, experimental methods, open data, open materials, preregistered

Received 5/19/17; Revision accepted 10/5/17

Replicability is a core value for empirical research, and there is increasing concern throughout psychology that more independent replication is necessary (Open Science Collaboration, 2015; Wagenmakers, Wetzels, Borsboom, van der Maas, & Kievit, 2012). Yet under the current incentive structure for science, replication is not typically valued for publication in top journals (Makel, Plucker, & Hegarty, 2012) or in metrics of research productivity (Koole & Lakens, 2012). One potential solution to this problem is to make replication an explicit part of pedagogy: that is, to teach students about experimental methods by asking them to run replication studies (Frank & Saxe, 2012; Grahe et al., 2012). Despite enthusiasm for this idea (Everett & Earp, 2015; King et al., 2016; LeBel, 2015; Standing, 2016), there is limited data beyond anecdotal reports and individual projects (e.g., Lakens, 2013; Phillips et al., 2015) to suggest that the pedagogical use of replications should be adopted on a broader scale.

In this article, we describe the pedagogical and methodological approach to replication research taken in our graduate-level experimental-methods course and address the practical barriers faced by instructors planning to incorporate replications into their courses. Students in our course conducted replications of published articles from the 2015 volume of the journal *Psychological Science*, and instructors rigorously reviewed their work at each major stage. The results of these replications are a microcosm of larger replication efforts, providing insight into both the difficulties of pedagogical replications and their promise as a method for improving the robustness of psychological research.

We assess the challenges facing a student in choosing an article of interest and—in a single attempt, within constraints of budget, expertise, and effort—reproducing the findings. We use a number of criteria for evaluating replication success, including statistical significance,

Corresponding Author:
Michael C. Frank, Stanford University–Psychology, 450 Serra Mall, Jordan Hall (Bldg. 420), Stanford, CA 94305
E-mail: mcfrank@stanford.edu

effect size, a Bayesian measure of evidence (Etz & Vandekerckhove, 2016), and subjective assessment with respect to the original authors' interpretations. Although each of these measures is imperfect, taken together they help provide a sense of the robustness and generalizability of the effects investigated, and perhaps more important, how easy it is for students to reproduce an effect to the degree that they could confidently build on it in their own future work.

We also describe in detail our process for conducting replications as part of classroom pedagogy. Although mentorship in experimental methods is an important part of the standard advising relationship in psychology, the classroom context allows for the elucidation of general principles of good research and discussion of how they can be modified to fit specific instances. And replication research in particular illustrates a number of important concepts—experimental design, power analysis, reporting standards, and preregistration, among others—more directly than open-ended projects, which require new conceptual development (see Frank & Saxe, 2012, for an extended argument).

There are significant limitations on what can be done in a single term, within the constraints of a course budget and the instructors' expertise. Nevertheless, were this approach implemented more widely, we believe the dividends—both scientific and educational—paid to the field as a whole would be considerable. We have made our course outline, project templates, and assignments available publicly at the Open Science Framework (https://osf.io/98ta4/files/). In addition, all of our most recent lecture slides and materials are available on our course Web site (http://psych254.stanford.edu). We hope that other instructors will share and reuse our materials as they consider how best to design a course customized to their aims and context.

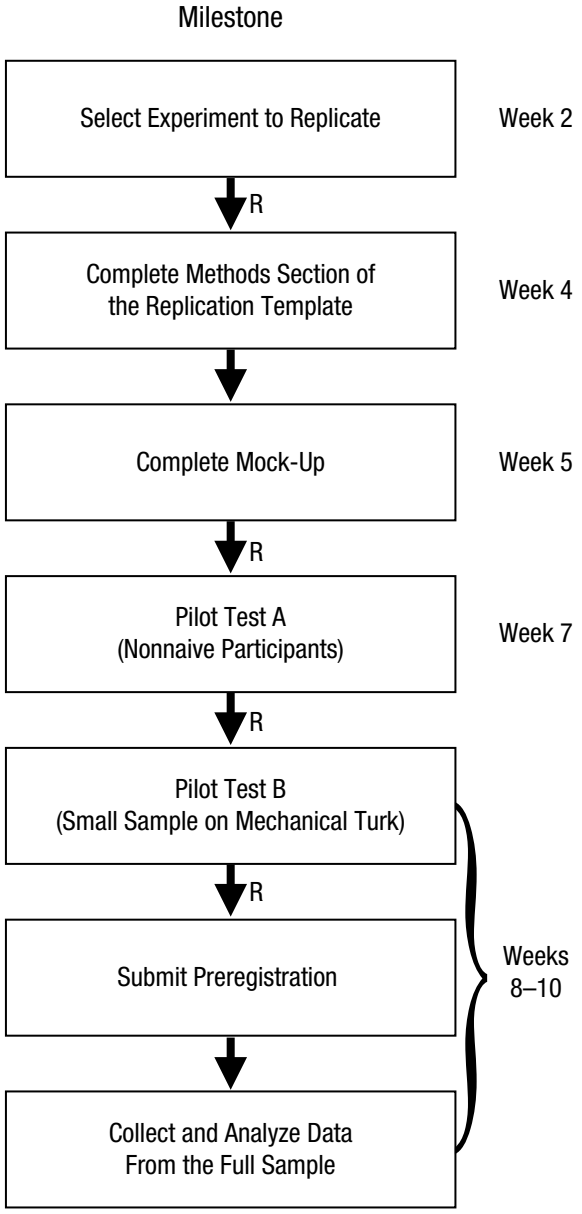

Fig. 1. Schematic of our class timeline for the replication studies. The *R*s indicate approximately when the instructor team reviewed students' materials. See the text for additional details.

Overview of the Project

These studies were completed as part of a graduate-level methodology course. At the beginning of the term, all students were told that they had the opportunity to contribute their individual class assignment to a group replication project. The requirements for joining the project were to conduct a preregistered replication of an experiment reported in the 2015 volume of *Psychological Science* and to contribute code, data, and materials to the write-up of the group project. A schematic of our class timeline is shown in Figure 1.

Eleven of the students' replication studies were included in the analyses reported here. Table 1 lists the original empirical studies on which these replications were based and their sample sizes. These original studies were from a wide variety of domains. One cluster of studies comprised investigations (some in applied contexts) of memory and visual attention; this cluster included work on capacity in mental rotation (Xu & Franconeri, 2015), memory for saved versus deleted computer files (Storm & Stone, 2015) and tracking of transitions between locations on smartphone apps (Liverence & Scholl, 2015). A second cluster consisted of studies of social perception and social judgment; these studies investigated judgments of faces and voices (Ko, Sadler, & Galinsky, 2015; Sofer, Dotsch, Wigboldus, & Todorov, 2015) and attributions of modesty, creativity, and expertise (Atir, Rosenzweig, & Dunning, 2015; Proudfoot, Kay, & Koval, 2015; Scopelliti, Loewenstein,

Table 1. Summary of the Original Studies and Replications

Original studies					Replication studies		
Study	Open data?	Open materials?	On MTurk?	*N*	Power standard	*N*	Instructor-rated fidelity
Atir, Rosenzweig, & Dunning, Experiment 1b	No	Yes	Yes	202	Other	50	6.67
Ko, Sadler, & Galinsky, Experiment 2	Yes	Some	No	40	Original	40	4.67
N. A. Lewis & Oyserman, Experiment 4	No	Yes	Yes	122	80% power	128	6.67
Liverence & Scholl, Experiment 1	No	Some	No	18	Original*	19	5.33
Proudfoot, Kay, & Koval, Experiment 1	No	No	Yes	80	80% power	84	6.67
Scopelliti, Loewenstein, & Vosgerau, Experiment 3	Yes	Yes	Yes	550	Other	124	5.67
Sofer, Dotsch, Wigboldus, & Todorov, Experiment 1	Yes	Yes	No	48	Other	95	4.33
Storm & Stone, Experiment 3	No	No	No	48	Original	61	4.00
Wang et al., Experiment 2	No	No	No	219	80% power	397	4.33
Xu & Franconeri, Experiment 1a	No	No	No	12	Other*	27	5.00
Zaval, Markowitz, & Weber, Experiment 1	Yes	Yes	Yes	312	Original	321	5.00

Note: All the original studies were reported in Volume 26 (2015) of *Psychological Science*. An asterisk indicates that the number of trials was modified. MTurk = Amazon Mechanical Turk.

& Vosgerau, 2015). The other studies investigated the effect of orientation to the future on retirement savings (N. A. Lewis & Oyserman, 2015), the effect of priming participants' future legacy on their proenvironmental behavior (Zaval, Markowitz, & Weber, 2015), and the effects of math anxiety on performance (Wang et al., 2015). Details of the individual studies chosen for replication are available in the Supplemental Material.

Disclosures

Preregistration

Our procedure for individual projects was to preregister the analytic script with tests of specific key hypotheses clearly marked. Links to the individual projects' preregistrations are available at https://osf.io/98ta4/wiki/. Prior to data collection for all the projects, we also preregistered the overall plan for the confirmation analyses reported in this article at https://osf.io/98ta4/.

Data, materials, and online resources

All code and data necessary to reproduce the analyses reported here are available at https://osf.io/98ta4/.

Measures

We report the method for determining sample size, all data exclusions, all manipulations, and all measures in all the replication studies.

Ethical approval

All the replication studies were approved by the Stanford University Institutional Review Board under Protocol 23274, "Reproducibility of Psychological Science and Instruction," and were conducted in accordance with the Declaration of Helsinki.

Method

All the replication studies were conducted on Amazon Mechanical Turk (MTurk). The students used JavaScript and HTML/CSS to reimplement their respective target studies in the Web browser, otherwise following the methods specified by the original authors as closely as possible. Of the 11 original studies included in the final sample, 64% had either some or all materials openly available (see Table 1). For 6 studies, students contacted the authors to request either materials or clarification of methods using a template e-mail that was customized by the students and reviewed by the instructors (this template is included in our course materials at the Open Science Framework). Responses were received in all but one case, all within a matter of days.

Participants

Individual sample sizes are given in Table 1. Each sample was recruited independently, using the same MTurk account but a different task title. Of the 11 original studies, 5 (45%) were conducted on MTurk (see Table 1). For the other 6 studies, demographic differences between the original sample and the replication sample (i.e., difference in age, sex, socioeconomic status, and, in some cases, national origin) are a factor in interpreting our findings (see the Supplemental Material for more discussion of this issue in individual cases).[1]

We determined sample sizes using a case-by-case selection criterion (see Table 1) intended to maximize the success of the projects while staying within the

constraints of our budget.[2] To ensure that they still met the desired criterion if some participants skipped through or did not complete the study, experimenters recruited 5% more participants than specified by their criterion. In three cases, we powered the replication attempts to 80% power, calculated using the original effect sizes. In four cases, we used the original sample size. In one case, the post hoc power analysis indicated that the original study was powered well above what would be required to find the effect of interest, so we chose a smaller sample size that yielded sufficiently high power to detect that effect. In another case, we increased the sample size from the original to compensate for reducing the number of trials per participant as part of the adaptation to MTurk, and in yet another, we reduced the number of trials but retained the original sample size, which resulted in sufficient power. In the last case, the original study used a multistep procedure in which the first step consisted of stimulus generation; we decreased the sample size in this first step and focused on achieving higher power in the second step.

Milestones, preregistration, and review

In addition to receiving as-needed guidance on their projects, the students advanced through a formal review process in which their work was inspected and critiqued several times by both the instructor and the teaching assistants (see Fig. 1). In the second week of the course, the students identified their replication targets and wrote brief proposals, which were reviewed for feasibility and concreteness. After their selection was approved, the students proceeded to write their Method section using the template developed by the Open Science Collaboration (2015) for replication reports. This section included a power analysis, the proposed sample size, and an explicit description of differences from the original study (see the course materials at https://osf.io/98ta4/files/ for the complete template). At the same time, the students worked on a mock-up of their experiment: a fully functional but unpolished outline, potentially including placeholders for unfinished elements of the design. Upon reviewing the completed Method section and mock-up, the teaching staff gave thorough feedback and suggested changes to be made before data collection began.

For each student's project, further reviews were triggered by the completion of data collection with two pilot samples. The first pilot test, Pilot Test A, involved a handful of nonnaive participants (e.g., the experimenter, other students). The goal of this pilot test was to ensure that all needed data were being accurately logged and that the analytic scripts for the confirmatory analyses functioned appropriately. After Pilot Test A was completed, the instructor or a teaching assistant critiqued and reviewed the student's experimental script, analytic code, and resulting output.

Once requested changes were made, the student conducted the second pilot test, Pilot Test B, using a handful of naive participants recruited from MTurk. The goal of this pilot test was to ensure that participants were able to complete the study and did not experience any substantial problems with the instructions or technical details of the study (all the students were instructed to give participants a way to leave free-form comments at the end of the study, prior to debriefing). At the conclusion of Pilot Test B, both the instructor and a teaching assistant reviewed the student's analytic code and its outputs for all confirmatory analyses. The goal of this code review was to ensure that all planned analyses were specified for the key test selected from the original article, and that all relevant exclusion criteria and manipulation checks specified by the original authors were implemented correctly.

After the review of Pilot Test B was completed and any requested changes had been made, the student was given authorization to recruit the prespecified number of MTurk participants and conduct the experiment. Prior to data collection, the students preregistered their replication reports, including the analytic script for all confirmatory analyses, using the Open Science Framework.

Statistical approach

Although the literature on meta-analysis and reproducibility generally recommends aggregating a single effect size of interest (Lipsey & Wilson, 2001; Open Science Collaboration, 2015), it was often challenging to identify a single key statistical test for a selected study. Usually, the original authors conducted a substantial number of statistical tests, either within one model (e.g., the authors reported both an interaction and main effect) or across multiple models (e.g., the authors reported results for three different dependent variables). Thus, we interpret the results for "key statistical tests" with caution (we discuss this concern further in the next section). Despite this interpretive difficulty, we attempted to use a single statistical test for each replication study.

Empirical Findings

All the reported results are from preregistered, confirmatory analyses. In two cases, the instructor team decided in its final review of the project (i.e., after data analysis) that the student's choice of key statistical test was not in fact a strong test of the original authors' primary theoretical claim. In each of these cases, the

original authors had fit regression models to the data, and there was not a single obvious test that corresponded directly to the authors' hypothesis. After discussion, the instructor team converged on a statistical test that they thought better corresponded to the original authors' intended hypothesis. We report results from these corrected tests here. We return to the issue of test selection later in this section, as we believe it is a critical theoretical issue in replication research (see, e.g., Monin, 2016, for discussion).

Subjective fidelity of the replications

Each member of the instructor team coded the fidelity of the replications, weighing whether the materials, sample, and task parameters differed from the original studies. Team members rated projects independently on a scale from 1 (loose replication with substantive deviations from the original) to 7 (close replication essentially identical to the original), and then discussed the ratings and made adjustments (without coming to full consensus). When the instructors disagreed, fidelity ratings were averaged across instructors. The mean rating was 5.29 (range: 4–6.67; see Table 1). Several studies were essentially identical to the originals, but others differed from the originals in population, number of trials, or method—all factors that might plausibly have had an effect on the results.

Subjective success of the replications

After data analysis, the subjective success of each replication (i.e., the extent to which it provided theoretical support for the original finding) was rated independently by the student who carried out the study and by each instructor. The rating scale had three response options (see the Supplemental Material for side-by-side plots of original and replication results): "no support" (coded as 0%), "partial support" (coded as 50%), and "full support" (i.e., the findings were consistent with the original interpretation; coded as 100%). The student's ratings agreed with the instructors' for 10 of the 11 replications. The average ratings given by instructors and students were 55% and 50%, respectively.

The modal replication study in our sample was judged to partially replicate the original findings; that is, some aspects of the observed findings were different from those reported by the original authors. Because our sample included only 11 studies, we did not attempt to conduct statistical tests (which would have been dramatically underpowered after correcting for multiple comparisons) and instead report general trends here. Overall, the replication studies had slightly higher success ratings if they were judged to be closer to the original. When the studies were split at the median subjective rating of fidelity, those with high fidelity received an average success rating of 70%, whereas those with low fidelity received an average success rating of 42%. Subjective success was also slightly greater for replications of original studies that were run on MTurk (run on MTurk: 60%; not run on MTurk: 50%) and for original studies that had open materials (materials publicly available: 57%; materials not publicly available: 50%), though we do not believe these differences are interpretable given our sample size.

Significance of the key effect

Of the 11 studies, 4 (36%) yielded a significant replication p value (i.e., $p < .05$).

Effect sizes and Bayes factors

To compute standardized effect sizes for the studies, we followed the recommendations of the Open Science Collaboration (2015) and converted test statistics to the correlation coefficient per degree of freedom, which is bounded between 0 and 1.[3] Although our preregistered analytic plan was to determine whether the reported effect size for each original study was included in the 95% confidence interval (CI) found in the replication, we instead opted for a straightforward comparison of point estimates. This choice was motivated by (a) the lack of consensus on how to properly calibrate the statistical details (e.g., a null distribution) of a test of inclusion within replication confidence intervals (Anderson et al., 2016; Gilbert, King, Pettigrew, & Wilson, 2016), (b) recent concerns that using confidence intervals in this manner may lead to misleading interpretations (Morey, Hoekstra, Rouder, Lee, & Wagenmakers, 2016), and (c) discrepancies in the original authors' reporting of effect sizes (which made computing confidence intervals difficult in some cases).

We first compared effect sizes from the original and replication studies (Fig. 2a). Effect sizes were highly correlated between the two sets of studies ($r = .73$, $p < .0001$), but they were generally smaller in the replications; on average, the effect size of the replications was 60% as large as the effect size of the original studies, 95% CI = [37%, 84%]. This pattern of results is consistent with findings from multistudy replication projects (e.g., Open Science Collaboration, 2015).

Next, following Etz and Vandekerckhove (2016), we compared Bayes factors between each original study and its replication study. We used the default test suggested by Rouder, Speckman, Sun, Morey, and Iverson (2009) and did not attempt to correct for publication bias.[4] We saw generally smaller Bayes factors for the

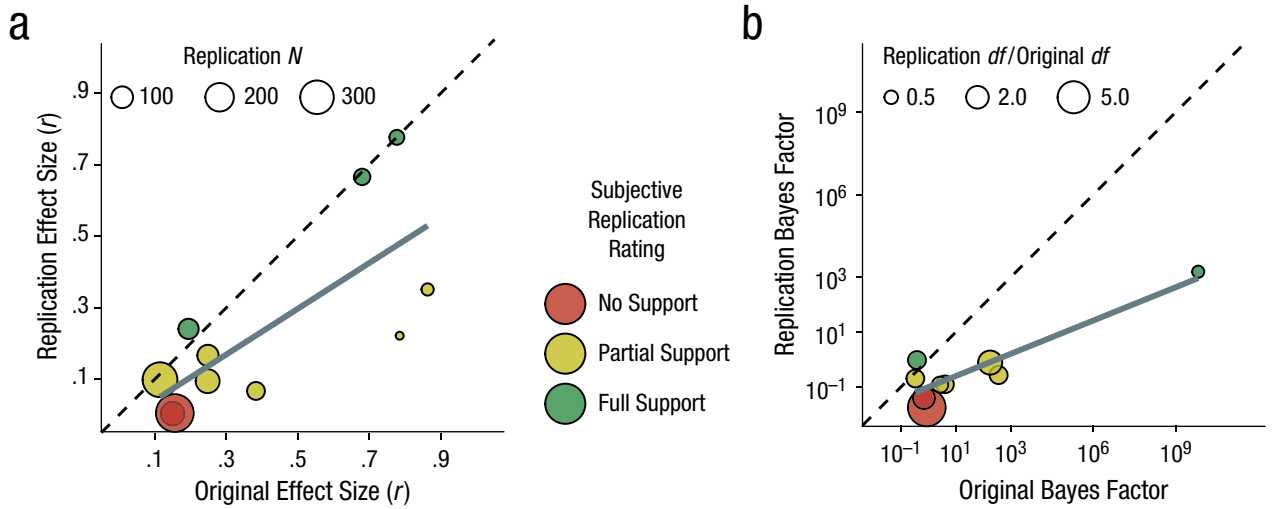

Fig. 2. Scatterplots showing the relation between the results of the replication and original studies. The graph in (a) shows results for effect size (*r*); the size of each plotted circle corresponds to the sample size in the replication study. The graph in (b) shows results for Bayes factors; the size of each plotted circle corresponds to the ratio of the test degrees of freedom in the replication study to the test degrees of freedom in the original study. In both plots, the color of the circles indicates the subjective assessment of replication success. In each graph, the solid line indicates the best-fitting linear regression line, and the dashed line represents unity (i.e., where the points would lie if the replications produced the same effect sizes or Bayes factors as the original studies). Note that the key statistic for one replication was a multivariate *F* test, so a default Bayes factor (see the text) could not be computed. Additionally, the original Bayes factor for one finding was many orders of magnitude greater than the others, so results for that replication are not displayed.

replication studies than for the original studies (Fig. 2b). Thus, in addition to yielding smaller effect sizes, the replications did not provide as strong support for the hypotheses of interest as the original studies did. Also, it appeared that the Bayes factors for the replication studies generally tracked nicely with subjective judgments of the replications' success. Finally, the Bayes factors for the effects in several of the original studies did not appear to show strong evidential value (e.g., Bayes factor < 3, indicating that the alternative hypothesis was less than 3 times more likely than the null), so it was not surprising that this was also the case in the replications of those original studies.

Implications for replication research

Statistical findings. From a purely statistical point of view, the results of our replication studies were underwhelming. Despite our relatively close adherence to the original protocols and relatively large sample sizes, the modal outcome was partial replication. In these cases of partial replication, some hint of the original pattern was observed, but often the key statistical test was not statistically significant, and the effect size was smaller and evidential value was lower than in the original study. We invite readers to browse the narrative descriptions of the individual replication attempts in our Supplemental Material to see the ways that patterns of empirical findings, beyond results of a single test, can differ between studies.

Our statistical findings mask a more optimistic message, however: For most of our studies, the next step for a motivated experimenter is clear. For example, we suspect that in some cases, a follow-up could find strong evidence for the phenomenon of interest by altering the particular planned analysis or titrating the difficulty of the stimulus materials. In other cases, the difficulty appeared to be statistical power, as differences were in the predicted direction and sometimes reached significance in subsidiary analyses; a follow-up would likely require a larger sample. And in still other cases, population differences suggest that follow-ups might be more successful if the stimuli or task were adapted. More generally, we believe that our work here underscores the importance of iterated replication for pinpointing empirical effects and refining theories (see, e.g., M. L. Lewis & Frank, 2016, for an example of this strategy and some discussion).

Many statistical factors that can reduce the success of replication attempts have been discussed in past work. These factors include analytic flexibility, context dependency, publication bias, and low statistical power, among others (e.g., Button et al., 2013; Ioannidis, 2005; Van Bavel, Mende-Siedlecki, Brady, & Reinero, 2016). For reasons of both design and scale, our study cannot disentangle the effects of these factors empirically. Nevertheless, the decreases in effect size and evidential value we report seem consistent with two explanations. First, our results likely reflect a "winner's curse"; that

is, initial publications may have tended to overestimate effects. Second, we also might have seen a reduction in effect size because the original studies were tailored to their specific contexts and populations, whereas our replications may not have been (Van Bavel et al., 2016; but cf. Inbar, 2016). It would be beneficial for future reports of empirical results to include a statement of constraints on generality so as to provide a guide to the conditions under which an effect is likely to be present (Simons, Shoda, & Lindsay, 2016).

Key tests. Our experiences performing and compiling the studies shed light on one further concern that we believe has been underreported in past studies (including Open Science Collaboration, 2015). The standard model of replication is based on the notion that a single key statistical test and its associated effect size are the properties of a study that can be targeted for replication. Yet, as we mentioned earlier, selecting this key test was difficult for most of the studies in our sample, and virtually impossible for some. We were often forced to consult other sections of the articles (e.g., abstracts, discussion sections) to gain clarity on what test was considered the critical one for the authors' interpretation. It is likely that some of our decisions would not be ratified by the original authors.

Indeed, no study we attempted to replicate was preregistered (though some were internal replications), and the general pattern of evidence across studies was often more important to the authors' conclusions than any particular test in any single study. In almost every study, the authors conducted multiple statistical tests, and the statistical models were often only tersely reported. In some cases, there was no conventionally reported and easily calculable effect-size measure (e.g., for mixed- and random-effect models; see Bakeman, 2005). These characteristics—which, in our experience, are endemic to published psychological work, including our own—make it extremely challenging to do replication research focused on the estimation of a single effect size.

Pedagogical Results

Pedagogical assessment

All of the 15 students in the class completed replication studies (2 with extensions going beyond the term). Of these 15, 1 chose a project outside of the sampling frame, and 3 opted out of the broader project prior to data collection.[5] The remaining 11 students contributed code, data, registrations, and materials to the final product. (In one case, a student did not complete the preregistration procedure correctly but still submitted a prespecified analysis plan to the instructors for review. We included this project in the final analysis.)

Because they were conducted by students within the constraints of a course, our studies had a number of limitations, including (a) limited time for iterative pilot testing and adjustment, (b) limited funding that precluded larger samples, and (c) limited domain expertise with respect to the specific effects that were chosen. But within these limitations, our group was able to produce a set of relatively close replication studies with generally good statistical power. Thus, our project shows what is possible for motivated students using freely available tools and resources.

Pedagogical implications

Our findings demonstrate that classroom replication projects are possible in one particular context. In the following subsections, we provide information that we hope will reduce barriers to initiating such projects and encourage more researchers and instructors to include replication studies as part of coursework in their courses. We discuss practical recommendations for implementation in a variety of classroom settings, key decisions that instructors must make, and common issues that arise during replication projects.

Models for students of different levels. Though other instructors' courses that incorporate replication projects will look quite different from our own, we recommend having the replication project as a centerpiece of the course and providing ample opportunity for feedback throughout the design phase. Without a central focus on the replication project, students would likely not have the time or motivation to complete a high-quality study; without multiple opportunities for extensive (and often interactive) feedback from instructors, the quality of the final studies will decline.

That said, depending on the student population and class constraints, there are many ways to include a replication project in a course. Table 2 presents three possible models of replication projects: for general undergraduate classes, advanced-undergraduate and early-graduate classes, and more advanced graduate classes. These models are not meant to be binding suggestions, but we hope that they inspire instructors to consider how replication research could be included in their own teaching practice. We next discuss some specific choice points.

Project selection. In our class, students selected projects on the basis of their own interest. We encouraged students with little programming background to opt for methodologically simple studies and more advanced students to

Table 2. Potential Models for Incorporating Replication Projects Into Classes of Different Levels

	Classroom model		
Course attribute	Model 1	Model 2	Model 3
Student group	Midlevel undergraduates	Advanced undergraduates or graduate students	Advanced graduate students
Learning goal	Gain experience with the research process	Gain research independence	Master best practices and tools
Group size	Full class or medium-size groups (7–8 students)	Small groups (2–4) or individuals	Individuals
Selection of original study	Single study chosen by the instructor or chosen by the student from a small curated list	Choice from a preselected list	Unconstrained student choice or choice within a meta-science sampling frame
Participant population	In-person convenience samples	University pool or Amazon Mechanical Turk	Amazon Mechanical Turk or targeted samples drawn from the original population
Workflow tools	GUI tools, code written by instructors	One key coding-based tool, GUI tools otherwise	Full ecosystem of key coding-based tools
Preregistration	Preregistration by the instructor	Preregistration drafted by the students	Preregistration by the students after instructor review
Dissemination	Study shared if a prespecified quality standard is met	Study shared if the instructor approves	Sharing is the default, but is subject to review

take on challenging designs. This level of freedom is not ideal for less advanced courses, and the set of available replication studies should likely be curated for students at lower levels. For example, in a large undergraduate class, it may be logistically simpler for the instructors to choose a small set of original studies that students can select from, or perhaps even pool resources and have the class collect multiple data sets targeting a single original study. Limiting the set of studies both gives instructors a greater degree of familiarity with the relevant literature and allows for greater scaffolding in developing experimental materials. A midlevel course might split the difference between a tightly curated list and complete student freedom, providing students with a broader—but still hand-selected—list of potential projects.

Participant population and sample size. For pedagogical replications to make a scientific contribution, they must be adequately powered and use participant populations comparable to those used in typical lab research. With these needs in mind, we chose to use MTurk for our replication studies. MTurk facilitates recruitment of diverse, high-quality samples (Buhrmester, Kwang, & Gosling, 2011; Crump, McDonnell, & Gureckis, 2013) that are large enough to enable replications of between-subjects designs (which typically require large samples). Because psychological research is increasingly conducted online, an additional pedagogical benefit of using MTurk is that it builds competence in relevant skills. Conducting a replication study on MTurk allows students to build a skill set that they can use throughout their academic careers.

Depending on the course design and total funds allocated to the course (if any), populations other than MTurk workers may be more feasible and appropriate. We recommend that instructors explore funding opportunities through teaching grants or departmental support, but note that a replication course without funding would be possible. For instance, community samples or university participant pools may provide high-quality data while reducing costs and eliminating a technical challenge in data collection.

Instructors must balance cost with the usefulness of replication research, given that larger sample sizes will result in more accurate estimates of the replication effect size (Button et al., 2013; Simonsohn, 2015). Sample planning based on analysis of the effect size in a previously published study is problematic because published effect sizes are likely inflated as a result of the winner's curse (Button et al., 2013; Hoenig & Heisey, 2001). But it may not be feasible to follow more conservative guidelines in all cases. For example, Simonsohn (2015) recommended collecting data from 2.5 times the original sample, which can be feasible for small or underpowered original studies, but may be unnecessary for studies that were large or adequately powered.

As a rule of thumb, for the type of replication research we are advocating, conducting a small number of high-powered studies using representative samples is preferable to conducting a large number of

underpowered studies in convenience samples. Results are much more likely to be interpretable by students and to make a contribution to the replication literature. We encourage instructors to discuss considerations of participant population and sample size with students for pedagogical purposes.

Workflow tools. All our studies were coded in JavaScript, HTML, and CSS, to be run in a standard Web browser. All analyses were written using R, a free and open-source platform for statistical analysis. The students wrote their final reports using R Markdown, a literate-programming environment in which statistical analysis code can be interspersed with text, figures, and tables. Using R Markdown as part of our workflow was extremely useful both for pedagogy and for encouraging reproducible research practices: At each phase of their projects, students reached a course milestone that required them to fill in additional sections of their own unified documents that would become their final reports. Writing in R Markdown allows students to easily share their cumulative work to date with instructors via a hyperlink, facilitating review of writing clarity, results, and code all in one location.[6]

Some of these open-source tools can feel inaccessible or intimidating to students with a limited programming background. Indeed, our students came from a variety of disciplinary backgrounds; many had relatively little experience with Web-based empirical studies or with incorporating code for reproducible data analysis directly into scientific reports. Yet by the end of the course, all gained sufficient proficiency with a suite of technical and conceptual tools that enabled them to carry out an independent project with support from their peers and the course staff. Our experience suggests that it is possible to convey key concepts in sufficient depth for students to learn to use a broad range of open-source and reproducible tools for their projects, even if mastery will require further experience.

Our explicit class goal was to provide students with experiences navigating the full ecosystem of open-source scientific tools. For classes at different levels, however, instructors will likely have different goals. It may be preferable for an intermediate course to focus on a single programming tool (e.g., R) and to use other GUI-based software for creating experiments, so as to minimize the learning burden for students. And for more introductory courses in research methods, instructors must gauge whether the added difficulty of a programming component is appropriate for the skills and backgrounds of their students.

Contact with the original authors. More than half of our students contacted the original authors to request study materials or clarification on methods or analysis, but for the purposes of our class, we chose not to require contact with the original authors. We strongly believe that methods and materials should be open, to allow for replication and productive science (Nosek et al., 2015), and hoped to empower the students to engage with and build on the published literature directly—without personal relationships or even personal contact. That is, we chose to have the students conduct good-faith replications given all information in the original articles and online supplemental materials. Authors were contacted only when there were issues about access to materials or ambiguities in design or analysis. Contacting original authors does have benefits, however. Authors can provide valuable feedback or guidance prior to data collection (albeit at the cost of some imposition on their time). In addition to being a professional courtesy, this practice can head off potentially taxing post hoc debates about methodological choices. In the end, it is up to individual instructors to decide what they believe is best for their students.

Ethical approvals. Ethical review standards vary across institutions and countries, but we suspect certain ethical concerns will commonly arise during attempts to implement class replication projects. First, instructors may be concerned that ethics approval is not possible within the time constraints of the class. To mitigate this issue, we suggest contacting ethics bodies well in advance of the term to determine whether to create a standard protocol that encompasses all individual studies (as we did), or to set a timeline for students and instructors to submit protocols for their studies very early in the course. Development of a protocol—whether individual or umbrella—will be simplified if the study or pool of studies to be replicated is preselected by instructors. Second, review boards may be concerned about risks particular to the collection of data by students (e.g., lack of debriefing or risk of breach of confidentiality because the experimenters are novices). We recommend building training in ethical issues into the syllabus of replication-based classes. This practice is positive for students and also mitigates potential concerns of ethical-approval boards.

General Discussion

We have reported on a series of 11 student replications of empirical studies published in the 2015 volume of *Psychological Science*. Our goal was to provide a proof of concept for pedagogical replications of recent findings in a top journal. Rather than attempting to gauge the truth of particular effects, we aimed to assess the challenges of replicating findings—selected on grounds of feasibility and interest—within the constraints of a course project. Such classes could become the backbone of future collaborative replication efforts (Everett

& Earp, 2015; Frank & Saxe, 2012), and could help provide a source of data on the robustness of the empirical literature more generally.

This vision of collaborative pedagogical replication requires instructors to share class results more broadly. We can envision a number of possible avenues for this kind of sharing. For example, if projects are shared openly in a repository such as the Open Science Framework (http://osf.io) or figshare (http://figshare.com) and tagged appropriately, they will be discoverable via search engines. Alternatively, a more structured method for sharing and discovery would be to upload files to specific replication curation Web sites, such as PsychFileDrawer (http://psychfiledrawer.org) or Curate Science (http://curatescience.org). Both of these models assume limited coordination across instructors, but more structured collaborations are possible. For example, the Collaborative Replications and Education Project (http://osf.io/wfc6u) selected a subsample of important and feasible studies for replication and helped provide materials to instructors with the explicit goal of encouraging cross-lab pedagogical replications. When they use these models, instructors must, of course, determine which projects will be released publicly, to ensure that disseminated replications meet prespecified standards. Instructors can incorporate dissemination into their course as a reward for high-quality work, as we did in the class project reported here.

In sum, our results demonstrate the practical possibility of performing replication research in the classroom. There are many challenges in attempting to ensure high-quality replications in a pedagogical setting—from checking experimental methods to reviewing analytic code for errors—but we argue that these are not just pedagogical challenges. They are challenges for psychological science. We believe that the openness and transparency we pursued in this project as part of our pedagogical goals should be models not just for other classes but for future research more broadly.

Action Editor

Daniel J. Simons served as action editor for this article.

Author Contributions

R. X. D. Hawkins, E. N. Smith, and M. C. Frank designed the project, supported data planning and data collection for all the individual studies, analyzed the data, and wrote the manuscript. M. C. Frank designed the course. C. Au, J. M. Arias, R. Catapano, E. Hermann, M. Keil, A. Lampinen, S. Raposo, J. Reynolds, S. Salehi, J. Salloum, and J. Tan planned and programmed the individual projects, collected the data, analyzed the data, and gave feedback on the manuscript. R. X. D. Hawkins and E. N. Smith contributed equally to this work and are listed alphabetically.

ORCID iD

Eric N. Smith https://orcid.org/0000-0001-5050-706X

Acknowledgments

We are grateful to the authors of the original studies, who provided materials and gave extensive comments on an earlier draft of this manuscript. An earlier draft of this manuscript was posted at https://osf.io/preprints/psyarxiv/p73he/.

Declaration of Conflicting Interests

The author(s) declared that there were no conflicts of interest with respect to the authorship or the publication of this article.

Funding

We thank the Department of Psychology and the Vice Provost for Graduate Education at Stanford University for providing funding to support the class. This work was partially supported by a National Science Foundation Graduate Research Fellowship under Grant No. DGE-1656518.

Supplemental Material

Additional supporting information can be found at http://journals.sagepub.com/doi/suppl/10.1177/2515245917740427

Open Practices

All data and materials, including all code necessary to reproduce the analyses reported here, have been made publicly available via the Open Science Framework and can be accessed at osf.io/98ta4/. The design and analysis plans for the individual replication studies and the confirmatory analyses reported in this article were preregistered at the Open Science Framework and can be also accessed at osf.io/98ta4/. The complete Open Practices Disclosure for this article can be found at http://journals.sagepub.com/doi/suppl/10.1177/2515245917740427. This article has received badges for Open Data, Open Materials, and Preregistration. More information about the Open Practices badges can be found at http://www.psychologicalscience.org/publications/badges.

Notes

1. Repeat administration of empirical work on MTurk is also a potential concern: Having a replication sample with nonnaive participants who previously participated in the original study or related research can lead to a number of unexpected effects (e.g., Chandler, Mueller, & Paolacci, 2014) or even reduced effect sizes (Chandler, Paolacci, Peer, Mueller, & Ratliff, 2015). Yet tracking participation in specific paradigms is an open challenge. We did not ask participants whether they had participated in similar research previously, as we suspected that asking this sort of question would lead to a large number of inaccurate responses due to participants' failure to distinguish between related experimental paradigms. In addition, half of

the MTurk population is estimated to change every 6 months (Stewart et al., 2015), and we expected that most of the original studies in our sample that had been conducted on MTurk had been performed at least a year previously.

2. We initially were allocated $1,500, but one student contributed personal research funds, so we had a total budget of approximately $1,700 for all the studies. Using estimates of study completion time from pilot testing, we attempted to set payment for our studies at approximately $6 per hour (Salehi et al., 2015).

3. For example, the conversion from a t statistic is given by $r = \sqrt{t^2/(t^2 + df)}$. Other formulas are given in Open Science Collaboration (2015, Supplemental Information).

4. This default test compares the null hypothesis of no effect against an alternative hypothesis, placing a Cauchy prior on effect size. The distribution of the prior has a mean, x_0, of 0 and scale parameter, γ, of 1; this distribution corresponds to a median absolute d of 1. A default Bayes factor for one original study could not be computed because of the statistical test that was used.

5. None of these 3 students expressed doubts about the success of their studies prior to opting out, and two of their studies were judged by the instructor team to have been successful replications. Thus, we have no evidence that these students opted out systematically because they believed that their study would not replicate the original results.

6. The current manuscript was written in this fashion as well; this writing method is also likely to reduce the frequency of statistical reporting errors (Nuijten, Hartgerink, van Assen, Epskamp, & Wicherts, 2015), given that errors are often introduced by transferring results between statistics and word-processing software.

References

Anderson, C. J., Bahnik, Š., Barnett-Cowan, M., Bosco, F. A., Chandler, J., Chartier, C. R., . . . Zuni, K. (2016). Response to comment on "Estimating the reproducibility of psychological science." *Science*, *351*, 1037.

Atir, S., Rosenzweig, E., & Dunning, D. (2015). When knowledge knows no bounds: Self-perceived expertise predicts claims of impossible knowledge. *Psychological Science*, *26*, 1295–1303.

Bakeman, R. (2005). Recommended effect size statistics for repeated measures designs. *Behavior Research Methods*, *37*, 379–384.

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

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munafò, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, *14*, 365–376.

Chandler, J., Mueller, P., & Paolacci, G. (2014). Nonnaïveté among Amazon Mechanical Turk workers: Consequences and solutions for behavioral researchers. *Behavior Research Methods*, *46*, 112–130.

Chandler, J., Paolacci, G., Peer, E., Mueller, P., & Ratliff, K. A. (2015). Using nonnaive participants can reduce effect sizes. *Psychological Science*, *26*, 1131–1139.

Crump, M. J. C., McDonnell, J. V., & Gureckis, T. M. (2013). Evaluating Amazon's Mechanical Turk as a tool for experimental behavioral research. *PLOS ONE*, *8*(3), Article e57410. doi:10.1371/journal.pone.0057410

Etz, A., & Vandekerckhove, J. (2016). A Bayesian perspective on the Reproducibility Project: Psychology. *PLOS ONE*, *11*(2), Article e0149794. doi:10.1371/journal.pone.0149794

Everett, J. A. C., & Earp, B. D. (2015). A tragedy of the (academic) commons: Interpreting the replication crisis in psychology as a social dilemma for early-career researchers. *Frontiers in Psychology*, *6*, Article 01152. doi:10.3389/fpsyg.2015.01152

Frank, M. C., & Saxe, R. (2012). Teaching replication. *Perspectives on Psychological Science*, *7*, 600–604.

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

Grahe, J. E., Reifman, A., Hermann, A. D., Walker, M., Oleson, K. C., Nario-Redmond, M., & Wiebe, R. P. (2012). Harnessing the undiscovered resource of student research projects. *Perspectives on Psychological Science*, *7*, 605–607.

Hoenig, J. M., & Heisey, D. M. (2001). The abuse of power. *The American Statistician*, *55*, 1–6.

Inbar, Y. (2016). Association between contextual dependence and replicability in psychology may be spurious. *Proceedings of the National Academy of Sciences, USA*, *113*, E4933–E4934.

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

King, M., Dablander, F., Jakob, L., Agan, M. L. F., Huber, F., Haslbeck, J. M. B., & Brecht, K. F. (2016). Registered reports for student research. *Journal of European Psychology Students*, *7*, 20–23.

Ko, S. J., Sadler, M. S., & Galinsky, A. D. (2015). The sound of power: Conveying and detecting hierarchical rank through voice. *Psychological Science*, *26*, 3–14.

Koole, S. L., & Lakens, D. (2012). Rewarding replications: A sure and simple way to improve psychological science. *Perspectives on Psychological Science*, *7*, 608–614.

Lakens, D. (2013). Using a smartphone to measure heart rate changes during relived happiness and anger. *IEEE Transactions on Affective Computing*, *4*, 238–241.

LeBel, E. (2015). A new replication norm for psychology. *Collabra*, *1*, Article 4. doi:10.1525/collabra.23

Lewis, M. L., & Frank, M. C. (2016). Understanding the effect of social context on learning: A replication of Xu and Tenenbaum (2007b). *Journal of Experimental Psychology: General*, *145*, e72–e80.

Lewis, N. A., & Oyserman, D. (2015). When does the future begin? Time metrics matter, connecting present and future selves. *Psychological Science 26*, 816–825.

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis*. Thousand Oaks, CA: Sage.

Liverence, B. M., & Scholl, B. J. (2015). Object persistence enhances spatial navigation: A case study in smartphone vision science. *Psychological Science*, *26*, 955–963.

Makel, M. C., Plucker, J. A., & Hegarty, B. (2012). Replications in psychology research: How often do they really occur? *Perspectives on Psychological Science*, *7*, 537–542.

Monin, B. (2016). Be careful what you wish for: Commentary on Ebersole et al. (2016). *Journal of Experimental Social Psychology*, *67*, 95–96.

Morey, R. D., Hoekstra, R., Rouder, J. N., Lee, M. D., & Wagenmakers, E.-J. (2016). The fallacy of placing confidence in confidence intervals. *Psychonomic Bulletin & Review*, *23*, 103–123.

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.

Nuijten, M. B., Hartgerink, C. H. J., van Assen, M. A. L. M., Epskamp, S., & Wicherts, J. M. (2016). The prevalence of statistical reporting errors in psychology (1985–2013). *Behavior Research Methods*, *48*, 1205–1226.

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

Phillips, J., Ong, D. C., Surtees, A. D. R., Xin, Y., Williams, S., Saxe, R., & Frank, M. C. (2015). A second look at automatic theory of mind: Reconsidering Kovács, Téglás, and Endress (2010). *Psychological Science*, *26*, 1353–1367.

Proudfoot, D., Kay, A. C., & Koval, C. Z. (2015). A gender bias in the attribution of creativity: Archival and experimental evidence for the perceived association between masculinity and creative thinking. *Psychological Science*, *26*, 1751–1761.

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian *t* tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, *16*, 225–237.

Salehi, N., Irani, L. C., Bernstein, M. S., Alkhatib, A., Ogbe, E., Milland, K., & Clickhappier. (2015). We are dynamo: Overcoming stalling and friction in collective action for crowd workers. In B. Begole (Ed.), *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (pp. 1621–1630). New York, NY: ACM.

Scopelliti, I., Loewenstein, G., & Vosgerau, J. (2015). You call it "self-exuberance"; I call it "bragging": Miscalibrated predictions of emotional responses to self-promotion. *Psychological Science*, *26*, 903–914.

Simons, D. J., Shoda, Y., & Lindsay, D. S. (2016). Constraints on generality (COG): A proposed addition to all empirical papers. *Perspectives on Psychological Science*, *12*, 1123–1128.

Simonsohn, U. (2015). Small telescopes: Detectability and the evaluation of replication results. *Psychological Science*, *26*, 559–569.

Sofer, C., Dotsch, R., Wigboldus, D., & Todorov, A. (2015). What is typical is good: The influence of face typicality on perceived trustworthiness. *Psychological Science*, *26*, 39–47.

Standing, L. G. (2016). How to use replication team projects in a research methods course. *Essays From X-Cellence in Teaching, XV*, 26–31.

Stewart, N., Ungemach, C., Harris, A. J. L., Bartels, D. M., Newell, B. R., Paolacci, G., & Chandler, J. (2015). The average laboratory samples a population of 7,300 Amazon Mechanical Turk workers. *Judgment and Decision Making*, *10*, 479–491.

Storm, B. C., & Stone, S. M. (2015). Saving-enhanced memory: The benefits of saving on the learning and remembering of new information. *Psychological Science*, *26*, 182–188.

Van Bavel, J. J., Mende-Siedlecki, P., Brady, W. J., & Reinero, D. A. (2016). Contextual sensitivity in scientific reproducibility. *Proceedings of the National Academy of Sciences, USA*, *113*, 6454–6459.

Wagenmakers, E.-J., Wetzels, R., Borsboom, D., van der Maas, H. L., & Kievit, R. A. (2012). An agenda for purely confirmatory research. *Perspectives on Psychological Science*, *7*, 632–638.

Wang, Z., Lukowski, S. L., Hart, S. A., Lyons, I. M., Thompson, L. A., Kovas, Y., . . . Petrill, S. A. (2015). Is math anxiety always bad for math learning? The role of math motivation. *Psychological Science*, *26*, 1863–1876.

Xu, Y., & Franconeri, S. L. (2015). Capacity for visual features in mental rotation. *Psychological Science*, *26*, 1241–1251.

Zaval, L., Markowitz, E. M., & Weber, E. U. (2015). How will I be remembered? Conserving the environment for the sake of one's legacy. *Psychological Science*, *26*, 231–236.