Assessing the Replicability of Psychological Science Through Pedagogy

Eric Smith*, Robert X. D. Hawkins*, the students of Psych 254, and Michael C. Frank

Department of Psychology, Stanford University

Author Note

These authors contributed equally. Thanks to the Stanford Department of Psychology and the Vice Provost for Graduate Education for funding to support the class.

Abstract

Replications are important to science, but who will do them? One proposal is that students can conduct direct replications as part of their training. As a proof-of-concept for this idea, here we report a series of 11 pre-registered, direct replications of findings from the 2015 volume of Psychological Science, all conducted as part of a graduate-level course. This work provides an estimate of the probability that a motivated graduate student can reproduce a previously published finding of interest on a first attempt. Congruent with previous findings, replication studies typically yielded smaller effects than originals: The modal outcome was partial support for the original claim. We describe the workflow and pedagogical methods that were used in the class and discuss challenges for adoption of this pedagogical model and replication research more broadly.

Keywords: Replication; Reproducibility; Pedagogy; Experimental Methods

Assessing the Replicability of Psychological Science Through Pedagogy

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, Maas, & Kievit, 2012). Under the current incentive structure for science, direct replication is not typically valued, however. One potential solution to this problem is to make direct 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, however (Everett & Earp, 2015; M. King et al., 2016; LeBel, 2015; Standing, 2016), there is limited data – beyond anecdotal reports or individual projects (Lakens, 2013; Phillips et al., 2015) – to support its efficacy.

In the current article, we provide evidence on this issue by reporting the results of replication projects conducted in a single class (a graduate-level experimental methods course). As part of the required work of the course, students conducted high-power, direct replications of published articles from the 2015 volume of the journal *Psychological Science*. These studies both give evidence about the state of replicability in the field and provide insight into the process of pedagogical replication.

It is a challenging proposition to estimate the replicability of an entire literature (cf. Open Science Collaboration, 2015). Making such an estimate minimally requires determining a sampling strategy, a notion of what constitutes a direct replication, and a standard for what constitutes a successful replication, among many other difficult decisions (Anderson et al., 2016; Gilbert, King, Pettigrew, & Wilson, 2016). In the current work, we focus on a quantity that is easier to estimate: the probability that a graduate student can choose an article of interest in *Psychological Science* and – in a single attempt, within constraints of budget, expertise, and effort – reproduce the basic

pattern of findings with sufficient fidelity to be able to build on it in future work. We believe that this notion of whether a piece of research supports cumulative progress is an important construct. Whether another independent scientist can build on a result is perhaps as important as – and in many cases easier to measure – whether the work is replicable in some more abstract sense.

With respect to pedagogy, we describe our process for conducting well-powered, high-quality, direct replications as part of classroom pedagogy. This process is dramatically facilitated by the use of standardized (and free) technical tools for statistical analysis, data and code sharing, and creation of web experiments. Nevertheless, there are significant limitations on what can be done in a single term and within the constraints of a course budget. We return in the Discussion to limitations on our approach.

In sum, we view the contribution of the current work to be two-fold. First, we provide a proof-of-concept and pedagogical model for conducting replications in the classroom. Second, we provide an initial estimate of what proportion of studies from the most recent volume of *Psychological Science* could be reproduced – at least as far as support for the same theoretical claims – by a motivated student.

Methods

We begin by describing the general method of the class, then discuss shared methods for the replications and the process of development for each project.

Class Outline

All projects were completed as part of a class (syllabus and materials available at http://psych254.stanford.edu). At the initiation of class, all students were told that

Citation	Expt	Open Data?	Open Materials?	Original on Turk?	N (orig)	N (rep)
Storm & Stone (2015)	3	No	No	No	48	61
Lewis & Oyserman (2015)	4	No	Yes	Yes	122	128
Scopelliti, Loewenstein, & Vosgerau (2015)	3	No	Yes	Yes	550	124
Liverence & Scholl (2015)	1	No	Some	No	18	19
Wang et al. (2015)	2	No	No	No	219	397
Sofer et al. (2015)	1	No	Yes	No	48	95
Ko, Sadler, & Galinksy (2015)	2	Yes	Some	No	40	40
Atir, Rosenzweig, & Dunning (2015)	1b	No	Yes	Yes	202	50
Proudfoot, Kay, & Koval (2015)	1	No	No	Yes	80	84
Zaval, Markowitz, & Weber (2015)	1	Yes	Yes	Yes	312	321
Xu & Franconeri (2015)	1a	No	No	No	12	27

Table 1

Summary of papers and experiments that were targets for replication in our projects.

they had the opportunity to participate in a group replication project to which they could contribute their class project. The requirements for joining the project were to conduct a pre-registered replication of a finding from the 2015 volume of *Psychological Science* and to contribute the code, data, and materials to the writeup.

Students selected projects based on interest, rather than via a systematic sampling strategy. Thus, all interpretation of our findings should be with respect to the distribution of student interest. Any reproducibility estimates arising from our analyses are estimates of the probability of reproducing a finding that is of interest, rather than findings in general. This sampling strategy may thus bias our results towards more exciting or novel findings, as well as those that were more feasible for students. A summary of the original findings is given in Table 1.

Participants

All study participants were recruited on Amazon Mechanical Turk; individual samples are given in Table 1.[1] All experiments were approved by the Stanford University institutional review board under protocol #23274, "Reproducibility of psychological science and instruction."

Materials and Methods

Of the 11 studies included in the final sample, 64% of studies had either some or all materials openly available. For 6 studies, however, students contacted the authors to request either materials or clarification of methods (using a template email that was customized by the students and reviewed by the instructors). Responses were received in all but one case, all within a matter of days.

Workflow

All materials and methods for our replication studies are available at https://github.com/StanfordPsych254. All experiments were coded in JavaScript, HTML, and CSS so as to be run in a standard web browser. All analyses were written using R, a free and open-source platform for statistical analysis. A schematic of our class workflow is shown in Figure 1.

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. This part of our workflow was extremely useful both pedagogically and for encouraging reproducible research practices. Using the template developed by Open

[1]Note that of the 11 studies included in the final sample, 5 (45%) were originally conducted on Mechanical Turk.

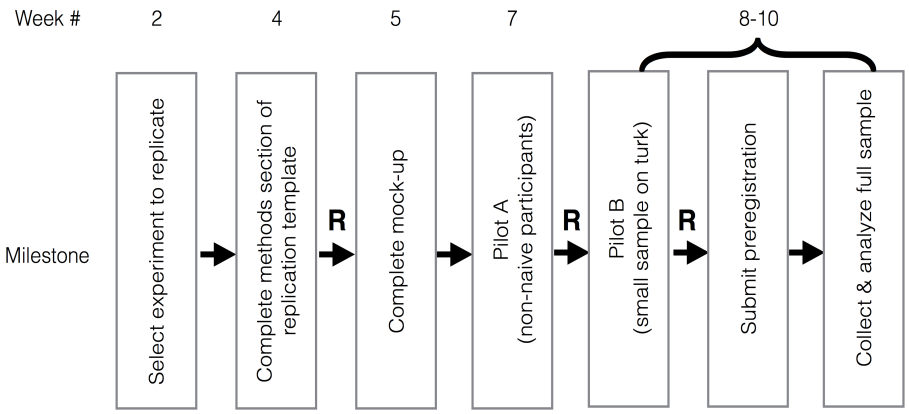

Figure 1. A schematic view of our class timeline for replication projects. R indicates the approximate timing of instructor team reviews of student materials.

Science Collaboration (2015) for replication reports, students created dynamic documents that included their proposed replication study, the code necessary to analyze their data, and the outputs of their data analysis (e.g. statistical tests, figures). These documents could be compiled into a variety of formats (e.g., HTML, PDF, Microsoft Word). This method of writing allows students to share a single compiled document via a hyperlink, facilitating review of writing, results, and code together in a single platform.[2] This writing method is also likely to reduce the frequency of statistical reporting errors (which appear to be regrettably common; Nuijten, Hartgerink, Assen, Epskamp, & Wicherts, 2015), given that many of these are likely introduced by copying and pasting between analysis outputs and manuscripts.

Preregistration and Review

In addition to as-needed guidance on projects, to ensure high quality replications, students went through a rigorous process in which each student's work was reviewed

[2]The current manuscript is written in this fashion as well.

several times by both the instructor (MCF) and the teaching assistants (ES and RXDH). Each student collected two pilot samples and review was coordinated with these. The first, "Pilot A," consisted of a handful of non-naive participants (e.g., the experimenter, other students). The goal of this pilot was to ensure that all needed data were being logged by the experiment script and that analytic scripts for the confirmatory analyses functioned appropriately. After Pilot A was completed, the instructor or a TA reviewed the student's experimental script (acting as a participant) and critiqued and reviewed the student's full analytic code and output.

Once requested changes were made, the student conducted "Pilot B," using a handful of naive participants recruited from Mechanical Turk. The goal of this pilot was to ensure that participants were able to complete the experiment and did not report any substantial issues with instructions or technical details of the experiment. (All students were instructed to give participants a way to leave free-form comments at the end of the experiment but prior to debriefing text). At the conclusion of Pilot B, both the instructor and a TA reviewed the student's analytic code and its outputs on the data for all confirmatory analyses. The goal of this code review was to ensure that all planned analyses were specified correctly and with sufficient detail to permit inclusion in the broader analysis.

After Pilot B review was completed, students were given authorization to collect the full sample of participants, contingent on having made any requested changes. Prior to data collection, the analytic script containing all confirmatory analyses was pre-registered using the Open Science Framework. In addition, prior to data collection, we pre-registered the confirmatory analyses reported in the current paper (`https://osf.io/rxz8m/`).

Sample size determination and statistical approach

After selecting a paper of interest, students chose a particular study from within that paper based on their judgment of interest and feasibility. The modal study chosen was the first, but several students chose later studies in a set as well.

Despite the general emphasis in meta-analytic and reproducibility efforts to aggregate a single effect size of interest (Lipsey & Wilson, 2001; Open Science Collaboration, 2015), it was often challenging to identify a single key statistical test in studies. Often a wide variety of statistical tests were conducted, often within one or several regression models with multiple specifications. Thus, we interpret "key statistical test" results with caution. Nevertheless, despite this interpretive difficulty, we attempted to use a single statistical test as our target for experiment planning and analysis.

Determining appropriate sample sizes is a second major challenge in replication research (Button et al., 2013; Simonsohn, 2015). Post-hoc power analyses are problematic because of the likelihood of inflation of effect sizes due to the "winner's curse" (Button et al., 2013; Hoenig & Heisey, 2001), but – especially within the constraints of a limited budget – it can be impossible to follow more conservative guidelines in all cases. For example, Simonsohn (2015) recommends 2.5x the original sample, which can be feasible for small or under-powered original studies but may lead to impractical or unnecessary recommendations for experiments that were initially large and/or adequately powered.

We eventually used a hybrid, case-by-case selection criterion that attempted to maximize the success of the project while staying within the constraints of our budget.[3]

[3]We initially were allocated $1,500 but one student contributed personal research funds, leading to a total cost of ≈ $1,700 for all studies. We attempted to set payment for our studies at approximately $6/hour based on timing estimated from Pilot A sessions.

In 4 cases, we powered the replication attempt to 80% post-hoc power. In 5 cases, we used the original sample size. In one case, the original experiment was extremely over-powered in the post-hoc analysis, so we chose a smaller sample size that still would yield high power. In one case, we reduced the number of trials as part of the adaptation to Mechanical Turk, and in one case of a multi-step procedure where the first step consisted of stimulus generation, we decreased the N in the first stage of the procedure (details available in SOM).

Individual Replications

Students chose findings from a wide variety of domains for replication. One cluster of studies included investigations of memory and visual attention (e.g., Xu & Franconeri, 2015), for example in applied contexts like remembering "deleted" files (Storm & Stone, 2015) or tracking transition between locations on smartphone apps (Liverence & Scholl, 2015). A second cluster came from studies of social perception and judgment, including judgments of faces and voices (Ko & Galinsky, 2015; Sofer, Dotsch, Wigboldus, & Todorov, 2015), as well as studies of attributions of modesty, creativity, and expertise (Atir, Rosenzweig, & Dunning, 2015; Proudfoot, Kay, & Koval, 2015; Scopelliti, Loewenstein, & Vosgerau, 2015). Other studies investigated increasing retirement savings by orienting participants to the future (N. A. Lewis & Oyserman, 2015), legacy priming (Zaval, Markowitz, & Weber, 2015), and the effects of math anxiety on performance (Wang et al., 2015). Details of individual replications are available in SOM.

Results

Pedagogical assessment

Of the 15 students in the class, all completed projects (two with extensions, one short and one more significant), but one failed to find a project of interest in the sampling frame and three opted out of the broader project prior to data collection.[4] The remaining 11 students contributed code, data, registrations and materials to the final product. (In one case, a student failed to complete the pre-registration procedure correctly but nevertheless submitted a pre-specified analysis plan to the instructors for review).

Each member of the instructor team coded the fidelity of replications, weighing whether the materials, sample, and task parameters differed from the original studies. Coded on a scale of 1 – 7 (with 1 being a loose replication with substantive deviations from the original, and 7 being essentially identical), the mean rating was 4.98. Several studies were essentially identical to the originals, but there was a group of others that included differences in population or number of trials that might have an effect on the results.

In two cases, in final review of the projects (after data collection), the instructor team decided that the student's choice of key statistical test was not in fact a strong test of the original authors' primary theoretical claim. In both of these cases (Proudfoot et al., 2015; Sofer et al., 2015), the original authors had fit a multiple regression model to the data and there was not a single obvious test that corresponded directly to the

[4]Note that these students did not express doubts about the success of their projects prior to opting out, and 2/3 of these projects were judged by the instructor team to have been successful replications. Thus we have no evidence that these students opted out systematically due to their belief that their project would fail.

authors' hypothesis. After discussion, the instructor team converged on a statistical test that they thought better corresponded to the original authors' intended hypothesis of interest. We report results from these corrected tests here. We return to this issue in the Discussion, as we believe that test selection is a critical theoretical issue in replication research.

Confirmatory analyses

Subjective judgments of replication. Students of the class and the instructors gave independent "replication ratings" to assess subjective judgement of the replication. Our rating system was based on the theoretical interpretation of the original finding, with 0 = no support for the original interpretation, .5 = some support for the original interpretation, 1 = full replication). We found that the ratings of the student carrying out the replication and the instructors were in agreement for 10 out of the 11 replications. The average rating given by instructors and students was 0.55 and 0.5 respectively.

Replicated		Original on Turk?		High fidelity?		Open Materials?		
		No	Yes	Yes	No	No	Some	Yes
Full	3	1	2	1	2	1	1	1
Partial	6	4	2	4	2	2	1	3
No	2	1	1	2	0	1	0	1
Total	11	6	5	7	4	4	2	5

Table 2

Counts of instructor team replication ratings, split by mediating variables.

The modal replication project in our sample was judged to be a "partial"

replication, meaning that some aspects of the observed findings were different from those reported by the original authors. While this conclusion is dispiriting, it is also perhaps unsurprising given the complexity of the analyses reported in many of the original papers. In addition, 5 of the tests were interactions, which tend to have lower statistical power (McClelland & Judd, 1993) and have fared poorly in previous replication studies (Open Science Collaboration, 2015).

The distribution of ratings are given in Table 2, along with breakdowns by various factors of interest. Since our sample only includes 11 studies, we describe general trends here rather than attempting to conduct statistical tests (which would be dramatically under-powered after correcting for multiple comparisons). Overall, projects had a somewhat larger probability of being judged successful if they were judged to be closer to the original: 75% vs. 43%. There were small numerical effects of being run on Mechanical Turk (60% vs. 50%) and having open materials (57% vs. 50%), though we do not believe these differences are interpretable given our sample size.

Significance of the key effect. We next turn to an assessment of the replication p-value to determine whether the key statistic found a significant effect at the traditional $p = .05$ level. Only 4 of the studies yielded a significant replication p-value. The well-documented "dance of the p-values" (Cumming, 2014) suggests that the inference drawn from a single replication p-value may not be informative, however, and other metrics such as effect sizes and Bayes factors may provide a more nuanced perspective.

Effect size and Bayes factor analysis. Effect sizes and Bayes factors improve upon inferences drawn from p-values in different ways. The effect size is an estimate of the *magnitude* of an effect, which is stable across sample sizes and allows for a more fine-grained comparison across studies than a binary significance criterion. The Bayes

Factor is an alternative hypothesis testing method that quantifies the *evidence* in favor of a hypothesis (Jeffreys, 1961). Among other advantages, it does not privilege the null hypothesis and can provide evidence either *for* or *against* (see also Scheibehenne, Jamil, & Wagenmakers, 2016; Wagenmakers, Morey, & Lee, 2016). While the Bayes Factor often agrees with the p-value as to which hypothesis is more likely (Wetzels et al., 2011), they often substantially disagree on the strength of the evidence.

These additional measures therefore give another sense in which replications can be successful: if we find roughly the same effect size or Bayes factor despite p-values on different sides of the $p = 0.05$ threshold, we should have increased confidence in the effect's reproducibility. To compute a standardized effect size, we followed Open Science Collaboration (2015) in converting test statistics to a measure of "correlation coefficient per df," which is bounded between 0 and 1[5]. We first compared effect sizes from the original and replication studies (Figure 2A). Effect sizes were highly correlated between the two sets of studies ($r = 0.73$, $p < .0001$), but they were generally smaller in the replications (60% of the original [37 - 84]).

While we preregistered an analysis determining whether the original reported effect size is included in the 95% confidence interval found in the replication, we instead opted for a straightforward comparison of point estimates. This choice was motivated by recent discussions of the utility and validity of interpreting confidence intervals in this manner (Morey, HoekstraLee, & Wagenmakers, 2016), as well as discrepencies in the reporting of effect sizes across the articles we included (which made computing confidence intervals difficult in some cases).

Next, following Etz & Vandekerckhove (2016), we compared Bayes Factors

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

between the original and replication. We used the default test suggested by Rouder, Speckman, Sun, Morey, & Iverson (2009) and did not attempt to correct for publication bias. This analysis revealed a number of interesting findings. First, as with effect sizes, we saw generally smaller Bayes Factors for the replications than the originals. Second, it appeared that replication Bayes Factors generally tracked nicely with subjective replication judgments as well. Finally, we note that the Bayes Factors for a number of the original effects did not appear to show strong evidential value (e.g. BF < 3); thus, it would have been somewhat surprising for our replications to show stronger evidence.

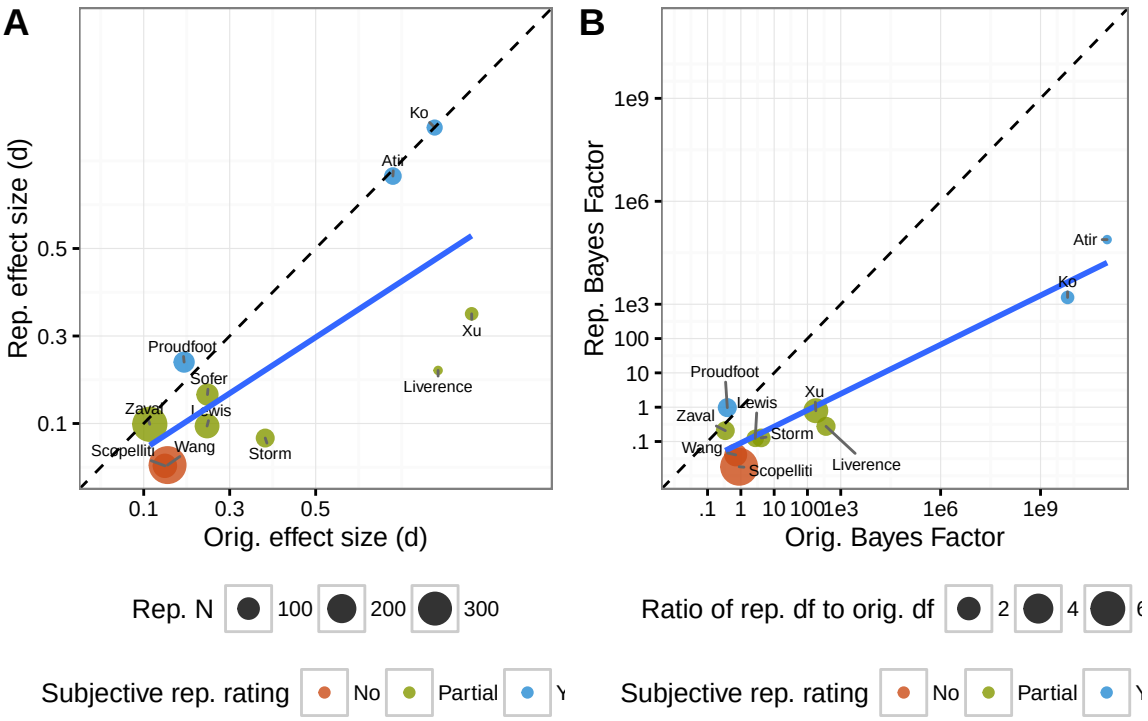

Figure 2. Replication effect size (A) and Bayes factor (B), plotted by the original effect size and Bayes factor, respectively. Point size shows replication N (A) and ratio of test degrees of freedom (B), color indicates subjective replication assessments by the authors. Note that the key statistic in the Sofer et al. (2015) replication was a multivariate F test, hence a default Bayes factor could not be computed.

Discussion

In this paper, we reported the results of a series of 11 direct replications of previously-published experiments from the 2015 volume of *Psychological Science*. We had two interlocking goals: first, to create a proof of concept for high-quality pedagogical replications, and second, to provide evidence on the ability of students replicate important, recent findings in a top journal. Despite our relatively tight adherence to the original experimental protocols and relatively large sample sizes, our results were disappointing. The modal outcome of our projects was partial replication, in which some hint of the original pattern was observed but the key statistical test was not statistically significant and showed both smaller effect size and lower evidential value than the same test in the original.

Our project aimed to assess the probability that a finding – selected on grounds of feasibility and interest – would be replicated with the time, money, and effort constraints of a course project. While these constraints might initially seem severe, our work here shows what is possible for motivated students using freely-available tools and resources. Indeed, the majority of the projects in our report were relatively close replications and included the same or greater numbers of participants. Nevertheless, they had a number of limitations, including at least: 1) limited time for iterative piloting and adjustment (M. L. Lewis & Frank, in press), 2) limits on available samples, and 3) limits on domain expertise with respect to the specific effects that were chosen. In addition, many papers published within our notional sampling frame (the 2015 volume of *Psychological Science*) were not feasible as projects, and students in our course were asked to take feasibility into account in choosing their projects as well as their own interests.

Our findings were congruent with the results of Open Science Collaboration (2015), revealing a relatively low rate of replicability by a number of tests. In addition,

in informal assessments of previous years' findings, we see a similar rate of replicability (Frank, n.d.). What are the causes of this relatively low rate?

Many statistical factors leading to lowered replicability have been discussed in past work, including analytic flexibility, publication bias, and low statistical power among others (e.g., Button et al., 2013; Ioannidis, 2005). On the other hand, an alternative viewpoint is that – especially for studies that leverage social constructs – exact replication may be difficult across contexts (Gilbert et al., 2016; Van Bavel, Mende-Siedlecki, Brady, & Reinero, 2016). For both reasons of design and scale, our study here cannot disentangle these factors empirically. Nevertheless, the consistent decreases in effect size and evidential value we report seem consistent with a "winner's curse" – that initial publications tend to over-estimate effects. However, in addition, although our sample was too small for statistical inference, numerically our results support replication fidelity in terms of sample and procedure as having some effects on success.

Our experiences performing and compiling the studies here sheds light on one further concern that we believe has been under-reported in past studies (including Open Science Collaboration, 2015). Our standard model of replication is based on the notion of a single statistical test – and its associated effect size – being the key properties of a study that can be targeted for replication. Yet, as we mentioned above, selecting this key statistical test was difficult for most of the projects in our sample, and virtually impossible for some. We were often forced to consult other sections of the papers (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, by the standards of clinical trials research, nearly every study we

replicated in this report would be classified as exploratory research. No protocol was pre-registered, and the general pattern of evidence across studies was often more important to the authors' conclusions than any particular test in any single study. Almost every study conducted multiple statistical tests, often associated with several statistical models with differing specifications. And in many cases, there was no conventionally-reported and easily calculable effect size measures (e.g., for mixed- and random-effect models, cf. Bakeman, 2005). This set of features – which, in our experience, are endemic to published psychological work – makes replication research within the conventional statistical paradigm extremely challenging.

In our view, two recommendations will improve this situation dramatically. First, pre-registration typically requires the selection of one or a small number of statistical analyses to be designated as critical for interpretation. Were this information available, it would have simplified the problem of selecting effects to replicate in several cases. Second, the widespread adoption of open data practices can dramatically facilitate the type of meta-research pursued here, in a number of ways. In addition to clarifying the (often terse) reporting in a paper, the availability of data facilitates the computation of the relevant meta-analytic variables.[6] No paper can perfectly predict the uses to which future investigators will put a piece of work; the availability of the source materials is a way of allowing more flexible reuse.

In conclusion, our results here demonstrate the practical possibility of performing replication research in the classroom. There are many challenges for ensuring high-quality replications in a pedagogical setting – from checking experimental methods

[6] We note that, although we were grateful that several papers in our sample shared data and materials, the use of open (non-proprietary) formats for archiving data and analytic code is often critical for other investigators attempting to reproduce analyses.

to reviewing analytic code for errors – but we would argue that these are not just pedagogical challenges. They are challenges for psychological science. We believe that the openness and transparency we pursued here as part of our pedagogical goals should be models not just for other classes but for future research more broadly.

Appendix: Project-By-Project Methods Addendum

For each project, we present a summary of the finding, the key statistical test, and the impressionistic outcomes of the replication. Figure 3 shows a side-by-side comparison of the key visualization for each study.

Storm & Stone (2015)

This paper found that when participants had the opportunity to save a file containing a list of target words before studying another file, they retained more information from the second study session. Our target for replication was Experiment 3, in which the effect was found to depend on the amount of information that was studied in the first file, with a greater effect for an eight-word list than a two-word list. The key statistical test we selected was the interaction in a 2x2 mixed-design ANOVA (save vs. no-save condition and two- vs. eight-word condition). The replication study failed to find evidence of the interaction, but ironically this was because we found evidence for a main effect of save condition: The authors' key manipulation succeeded for both load conditions. Perhaps even two words was sufficient to create interference for our online sample.

Lewis & Oyserman (2015)

This paper investigated motivations to save for retirement, and specifically whether changing participants' relationships to the future (making it feel more imminent) would

lead them to report that they would begin saving for retirement sooner. Our target was Study 4, which found that seeing the time to retirement in days (10,950) rather than years (30) caused participants to say they would begin saving sooner, even controlling for age, education, and income in a linear model (there was also a manipulation of whether saving was incremental that did not result in an effect). We failed to find a significant relationship between the time metric manipulation and saving time. However, when we did not include demographic controls, we did find a marginally-significant relationship. In exploratory analyses, we also observed an un-predicted effect of income such that participants who reported higher income selected to start saving sooner.

Scopelliti, Lowenstein & Vosgerau (2015)

The goal of this paper was to examine whether people are accurately calibrated in their estimates of how others will respond to their attempts at self-promotion. Our target was Experiment 3, which showed that participants, who were instructed to describe themselves in such a way that others would like them, expected to be liked more – but were actually liked less – than participants who were asked to describe themselves but not additionally instructed to maximize others' interest in meeting them. The authors performed a series of regressions on four dependent variables capturing judgments about profile writers (interest, liking, bragging, and success), analyzing these as a function of condition (control vs. "maximize interest") and who the evaluator was (the writer vs. an independent sample). Power analysis for this design was quite complex because the procedure had two stages: in a first stage participants generated descriptions and predicted judgments, while in a second stage a separate sample rated the descriptions.

We were faced with a choice: either we could have re-rated the descriptions

gathered in the original study (by contacting the author and treating these as "materials" for the study), effectively replicating only stage two, or we could redo the entire two-step procedure. We elected to conduct a replication of the full procedure, but in our replication to decrease cost we focused on the liking judgments, which appeared to be central to the authors' conclusions. Our key test was the interaction of rater (writer vs. independent) and condition.[7] However, to power even this smaller study adequately was outside of our budget, so instead of having each judge rate a subset of 10 out of 100 total profiles, we collected a smaller sample of profiles (18) and had every profile rated by each judge. We failed to find the predicted interaction in this new sample. Instead, we found a main effect such that stage one participants predicted that raters would like them more than the stage two participants actually did. Despite our relatively high statistical power in the main test, however, we may have suffered from low power at the item level due to the relatively small number of profiles we collected in the replication study.

Liverence & Scholl (2015)

This paper tested the theory that persistent object representations could assist in spatial navigation, using a navigation paradigm in which participants used key presses to move though a grid of pictures that were visible only one at a time. In Experiment 1, our target, found that participants located targets faster when navigation involved persistence cues (via sliding animations) than when persistence was disrupted (via temporally matched fading animations). The key test for our replication was a simple *t*-test comparing speed to find the target in the two conditions (slide vs. fade), not

[7]We note that this project was submitted late and the student failed to preregister this analysis formally; however, the pre-written analysis protocol went through pre-data collection review by MCF.

controlling for learning across the duration of the experiment. We failed to find this overall difference. However, in an exploratory followup analysis using a linear mixed effect model, we did find a highly-significant effect of condition when controlling for learning epoch. We speculate that the very large effect found by the original authors may have led them to select a simpler analysis that did not control for learning effects. Our data – which were more variable – required this control.

Wang et al. (2015)

This paper reported that the relationship between math anxiety and math performance follows an inverted-U pattern among college undergraduate students with higher intrinsic math motivation, whereas this same relationship is linearly negative among students with lower intrinsic math motivation. We targeted Study 2, the authors' own replication study, in which participants filled out three surveys about their math motivation and anxiety and completed a math task to measure performance. The key effect we targeted was the interaction of math anxiety squared and math motivation in predicting performance. We failed to find evidence for this effect. In contrast to the original study, we found a simple linear trend such that performance decreased with increasing math anxiety for both high and low math-motivation groups.

Sofer et al. (2015)

This paper tested the hypothesis that face typicality affects trustworthiness judgments such that the most typical face will be judged to be the most trustworthy. In Study 1, our target, faces of varying typicality were rated on their trustworthiness and attractiveness; while trustworthiness peaked at the most typical face, less typical faces were judged more attractive. While the original class projects selected a global,

item-wise regression model as the key statistical test of interest, the instructor team (post-data collection) determined that the fit of this model did not in fact correspond most closely with the authors' stated hypothesis. We thus selected the interaction between face typicality and rating condition as our key test statistic and used the more standard by-participant regression model (which the original paper also reported subsequent to the by-items model) as the target model. Our replication study successfully showed this same interaction.[8] We note that, despite the significant interaction, in our study trustworthiness judgments did not change substantially with typicality (though attractiveness did) and trustworthiness did not peak numerically at the most typical face. Thus, despite replicating one key statistical test, our results are at odds with some of the original paper's claims.

Ko, Sadler, & Galinksy (2015)

This paper investigated vocal cues to social hierarchy. Our target for replication was Experiment 2, which assessed whether participants used particular acoustic cues to make inferences about speakers' level of social hierarchy, particularly that participants from the high-rank condition in Experiment 1 were rated as more highly ranking. This paper employed a two-stage procedure in which participant-generated materials were rated by an independent sample. However, in contrast to the study of Scopelliti, Lowenstein & Vosgerau (2015), because materials were collected in Experiment 1 and were available openly, we elected to replicate only Experiment 2, the judgment study. The key test statistic was the main effect of hierarchy condition on "behavior score" (an index of whether speakers would engage in high-status behaviors). The authors were

[8] We were unable to reproduce the numerical results from this model using the data from the original study, but we report the (larger) test statistics from our recomputed version.

also interested in which particular vocal cues predicted participants judgments, but we judged these analyses to be more descriptive and exploratory. We successfully replicated the main effect of condition on hierarchy judgment, and additionally found a comparable main effect of speaker gender.

Atir, Rosenzweig, & Dunning (2015)

This paper investigated how perceived knowledge affects "overclaiming" (claiming knowledge that you do not have). We targeted Study 1b, which asked participants to rate their knowledge about finances, and then asked them to rate their knowledge about a set of financial terms, some of which were invented (and hence for which a positive knowledge statement would necessarily be an overclaim). The original study found that participants with higher self-rated financial knowledge also overclaimed at a higher rate. The key test statistic for this study was the coefficient in a regression model predicting overclaiming as a function of claimed personal financial knowledge, controlling for accuracy in the true financial knowledge questions. We successfully replicated this effect, despite our use of a much smaller sample (due to the extremely high post-hoc power of the original study).

Proudfoot, Kay, & Koval (2015)

This paper examined whether judgments of creativity are influenced by the gender of the creator. In Study 1, they tested the hypothesis that "out of the box" creativity was more associated with masculine traits than feminine traits by manipulating the definition of creativity that participants saw (convergent vs. divergent) and asking participants to judge the centrality of different traits to that definition. The original pre-registration selected a subsidiary analysis of a main effect as the key test statistic,

but in post-data collection review the instructor team decided that a test of the interaction between trait gender and definition condition was closer to the central theoretical claim of the paper; we use this latter test as the key test statistic. Our replication study found an significant interaction of similar magnitude.

Zaval, Markowitz, & Weber (2015)

This paper investigated whether encouraging participants to think about their personal legacy could increase concern for the environment. The main study of the paper investigated whether a legacy priming writing exercise would increase a variety of environmental measures, including donations to an environmental charity, pro-environmental intentions, and climate change beliefs. The key test we selected was the effect of condition (priming vs. control) on behavioral intentions in an analysis of variance. We observed a marginally significant effect in the same direction in our replication, as well as support for the mediating relationship of legacy motives on behavioral intentions. Our study may have been under-powered due to budgetary constraints; we ran a sample of comparable size to the original sample of 312 participants. We also may have observed a smaller effect size due to the relatively smaller amount of time our participants spent on the legacy prime writing exercise.

Xu & Franconeri (2015)

This paper explored constraints on visual working memory in the context of mental rotation. In Experiment 1a, participants performed a verbal suppression task to control for use of verbal encoding while attempting to remember a cross with four colored bars, which sometimes switched colors; performing mental rotation significantly impaired ability to detect these bar swaps. To port the task to an online framework we decreased

the number of trials and increased the number of participants we tested. The key test statistic was a comparison of K (a metric of the number of visual features remembered) between the rotation and no-rotation conditions. We observed a marginally-significant effect in our replication study. Our study showed a floor effect on capacity such that participants remembered on average very little; it is possible that small details in the displays (or the differing sample available online) led to this floor effect.

References

Anderson, C. J., Bahnik, S., Barnett-Cowan, M., Bosco, F. A., Chandler, J., Chartier, C. R., ... others. (2016). Response to comment on "Estimating the reproducibility of psychological science". *Science*, *351*(6277), 1037–1037.

Atir, S., Rosenzweig, E., & Dunning, D. (2015). When knowledge knows no bounds: Self-perceived expertise predicts claims of impossible knowledge. *Psychological Science*, *26*(8), 1295–1303.

Bakeman, R. (2005). Recommended effect size statistics for repeated measures designs. *Behavior Research Methods*, *37*(3), 379–384.

Button, K. S., Ioannidis, J. P., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S., & Munafo, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, *14*(5), 365–376.

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science*, *25*, 7–29.

Etz, A., & Vandekerckhove, J. (2016). A Bayesian perspective on the reproducibility project: Psychology. *PloS One*, *11*(2), e0149794.

Everett, J. A., & 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*.

Frank, M. C. (n.d.). Assessing p(replication) in a practical setting. Retrieved from `http://babieslearninglanguage.blogspot.com/2015/03/estimating-preplication-in-practical.html`

Frank, M. C., & Saxe, R. (2012). Teaching replication. *Perspectives on Psychological Science*, *7*(6), 600–604.

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

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*(6), 605–607.

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

Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS Medicine*, *2*(8), e124.

Jeffreys, H. (1961). Theory of probability. Oxford: Clarendon Press.

King, M., Dablander, F., Jakob, L., Agan, M., Huber, F., Haslbeck, J., & Brecht, K. (2016). Registered reports for student research. *Journal of European Psychology Students*, *7*(1).

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

Lakens, D. (2013). Using a smartphone to measure heart rate changes during relived happiness and anger. *Affective Computing*, *4*(2), 238–241.

LeBel, E. (2015). A new replication norm for psychology. *Collabra*, *1*(1).

Lewis, M. L., & Frank, M. C. (in press). Understanding the effect of social context

on learning: A replication of Xu and Tenenbaum (2007b). *Journal of Experimental Psychology: General*.

Lewis, N. A., & Oyserman, D. (2015). When does the future begin? Time metrics matter, connecting present and future selves. *Psychological Science*, 0956797615572231.

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

Liverence, B. M., & Scholl, B. J. (2015). Object persistence enhances spatial navigation a case study in smartphone vision science. *Psychological Science*, 0956797614547705.

McClelland, G. H., & Judd, C. M. (1993). Statistical difficulties of detecting interactions and moderator effects. *Psychological Bulletin*, *114*(2), 376.

Morey, R. D., Hoekstra, R., Jeffrey N, Lee, M. D., & Wagenmakers, E.-J. (2016). The fallacy of placing confidence in confidence intervals. *Psychonomic Bulletin & Review*, *23*(1), 103–123.

Nuijten, M. B., Hartgerink, C. H., Assen, M. A., Epskamp, S., & Wicherts, J. M. (2015). The prevalence of statistical reporting errors in psychology (1985–2013). *Behavior Research Methods*, 1–22.

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

Phillips, J., Ong, D. C., Surtees, A. D., Xin, Y., Williams, S., Saxe, R., & Frank, M. C. (2015). A second look at automatic theory of mind: Reconsidering Kovacs, Teglas, and Endress (2010). *Psychological Science*, 0956797614558717.

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*, 0956797615598739.

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*(2), 225–237.

Scheibehenne, B., Jamil, T., & Wagenmakers, E.-J. (2016). Bayesian evidence synthesis can reconcile seemingly inconsistent results: The case of hotel towel reuse. *Psychological Science*, *27*(7), 1043–1046. Retrieved from `http://pss.sagepub.com/content/27/7/1043.short`

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*(6), 903–914.

Simonsohn, U. (2015). Small telescopes detectability and the evaluation of replication results. *Psychological Science*, 0956797614567341.

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.

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*(2), 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*, *113*(23), 6454–6459.

Wagenmakers, E.-J., Morey, R. D., & Lee, M. D. (2016). Bayesian benefits for the

pragmatic researcher. *Current Directions in Psychological Science*, *25*(3), 169–176. http://doi.org/10.1177/0963721416643289

Wagenmakers, E.-J., Wetzels, R., Borsboom, D., Maas, H. L. van der, & Kievit, R. A. (2012). An agenda for purely confirmatory research. *Perspectives on Psychological Science*, *7*(6), 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*(12), 1863–1876.

Wetzels, R., Matzke, D., Lee, M. D., Rouder, J. N., Iverson, G. J., & Wagenmakers, E.-J. (2011). Statistical evidence in experimental psychology an empirical comparison using 855 t tests. *Perspectives on Psychological Science*, *6*(3), 291–298.

Xu, Y., & Franconeri, S. L. (2015). Capacity for visual features in mental rotation. *Psychological Science*, *26*(8), 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*(2), 231–236.

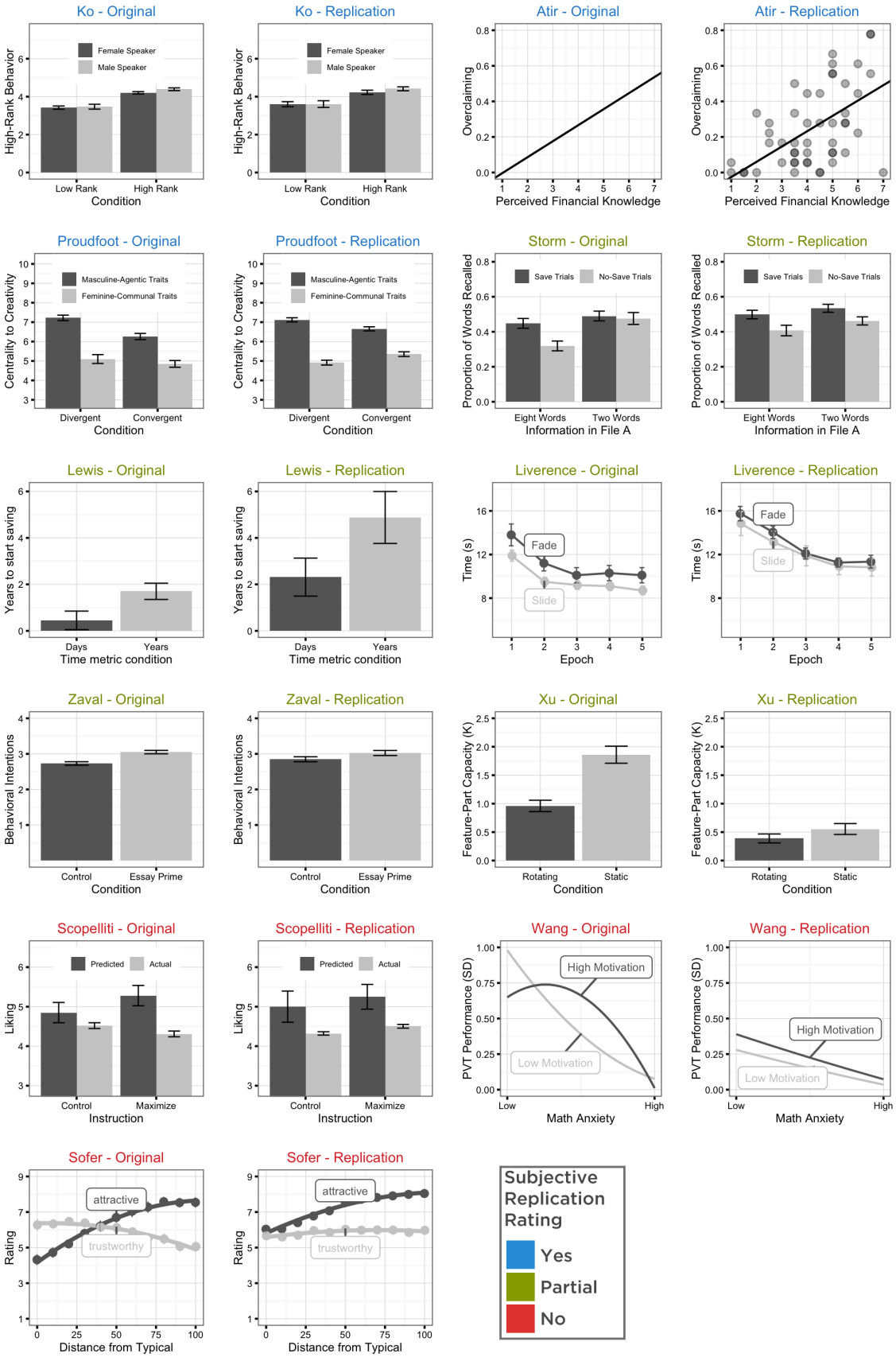

Figure 3. Side-by-side plots for each attempted replication. Error bars show standard error of the mean. Original data estimated from figures when not otherwise available. Replications are sorted and color coded by subjective replication success.