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

Author Note

*These authors contributed equally and are listed alphabetically.

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 pre-registered replications of findings from the 2015 volume of Psychological Science, all conducted as part of a graduate-level course. Congruent with previous studies, replications typically yielded smaller effects than originals: The modal outcome was partial support for the original claim. This work documents the challenges facing motivated students in reproducing 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

Improving 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). Yet under the current incentive structure for science, replication is not typically valued for publication (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; M. King et al., 2016; LeBel, 2015; Standing, 2016), there is limited data beyond anecdotal reports or individual projects (D. Lakens, 2013; Phillips et al., 2015) to support its efficacy in producing wide-scale pedagogical adoption.

In the current article, we address the practical barriers of completing replications as part of required coursework and discuss the methodological decisions that such replication projects must address more broadly. Towards this aim, we report the results of replication projects conducted in a graduate-level experimental methods course. Students conducted replications of published articles from the 2015 volume of the journal *Psychological Science*. These studies provide 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 consider a number of criteria for evaluating replication success, including statistical significance, effect size, a Bayesian measure of evidence (Etz &

Vandekerckhove, 2016), and a subjective assessment with respect to the original authors' interpretations. While each of these is imperfect, taken together these measures suggest that replications performed in the classroom are representative of larger, more systematic efforts. Perhaps more importantly, these results provide a sense of how easy it is for a student to reproduce an effect to the degree that they could confidently build on it in their own future work.

We also describe our process for conducting replications as part of classroom pedagogy. Although mentorship in experimental methods is an important part of the standard advising relationship, the classroom context allows for 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 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 paid to the field as a whole – both scientific and educational – would be considerable.

We begin by providing the details of our course projects to give a sense of what a classroom replication project entails. We then discuss broader decision points for the course design that other instructors might consider. Lastly, we report the results our replications, highlighting the importance of wide-scale replication efforts.

To encourage others to share our materials, our course outline, project templates, and assignments are available publicly at `https://osf.io/98ta4/files/`. In addition, all of our most recent lecture slides and materials are available on our course website at

http://psych254.stanford.edu. We hope that others will share and reuse our materials as they consider how best to design a course customized to their context.

Citation (Psychological Science; 2015)	Expt.	Original Study: Open Data?	Original Study: Open Materials?	Original Study: On MTurk?	Original Study: N (orig)	Replication Study: Power Standard	Replication Study: N (rep)	Replication Study: Instructor Fidelity
Atir, Rosenzweig, & Dunning	1b	No	Yes	Yes	202	Other	50	6.67
Ko, Sadler, & Galinksy	2	Yes	Some	No	40	Original	40	4.67
Lewis & Oyserman	4	No	Yes	Yes	122	80% power	128	6.67
Liverence & Scholl	1	No	Some	No	18	Original*	19	5.33
Proudfoot, Kay, & Koval	1	No	No	Yes	80	80% power	84	6.67
Scopelliti, Loewenstein, & Vosgerau	3	Yes	Yes	Yes	550	Other	124	5.67
Sofer et al.	1	Yes	Yes	No	48	Other	95	4.33
Storm & Stone	3	No	No	No	48	Original	61	4.00
Wang et al.	2	No	No	No	219	80% power	397	4.33
Xu & Franconeri	1a	No	No	No	12	Other*	27	5.00
Zaval, Markowitz, & Weber	1	Yes	Yes	Yes	312	Original	321	5.00

Table 1

*Summary characteristics of original studies and our replications. All project materials available publicly at osf.io/98ta4/files/, and links to individual project preregistrations, reports, and web experiments are available at osf.io/98ta4/wiki/. * marks projects where the number of trials was modified.*

Disclosures

Preregistration

As described below, our procedure for individual project preregistrations was to register the analytic script with specific key hypothesis tests clearly marked. Individual project preregistration links are given in Table 1. Prior to data collection for all projects, we also pre-registered the confirmatory analyses reported in the current paper at `https://osf.io/rxz8m/`.

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 how we determined our sample size, all data exclusions, all manipulations, and all measures in the study (see Simmons, Nelson, & Simonsohn (2011)).

Subjects

All replications were approved by the Stanford University Institutional Review Board under protocol #23274, "Reproducibility of psychological science and instruction" and was conducted in accordance with the Declaration of Helsinki.

Conflicts of Interest

The authors declare that they have no conflicts of interest with respect to the authorship or the publication of this article.

Author Contributions

RXDH, ENS, and MCF designed the project, supported data planning and data collection for all projects, analyzed the data, and wrote the paper. CA, JMA, RC, EH, MK, AL, SR, JR, SS, JS, and JT planned individual projects, programmed studies, collected data, analyzed data, and gave feedback on the paper.

Acknowledgements

Thanks to the Stanford Department of Psychology and the Vice Provost for Graduate Education for funding to support the class. We are grateful to the authors of the original studies who provided materials and gave extensive comments on an earlier draft of this manuscript.

Prior Versions

An earlier draft of this manuscript was posted at `https://osf.io/preprints/psyarxiv/p73he/`.

Methods

All projects were completed as part of a graduate-level methodology class. At the initiation of class, 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 pre-registered replication of a finding from the 2015 volume of *Psychological Science* and to contribute the code, data, and materials to the writeup. A schematic of our class timeline is shown in Figure 1.

A summary of the original empirical studies chosen for replication is given in Table 1. Students chose findings from a wide variety of domains for replication. One cluster of

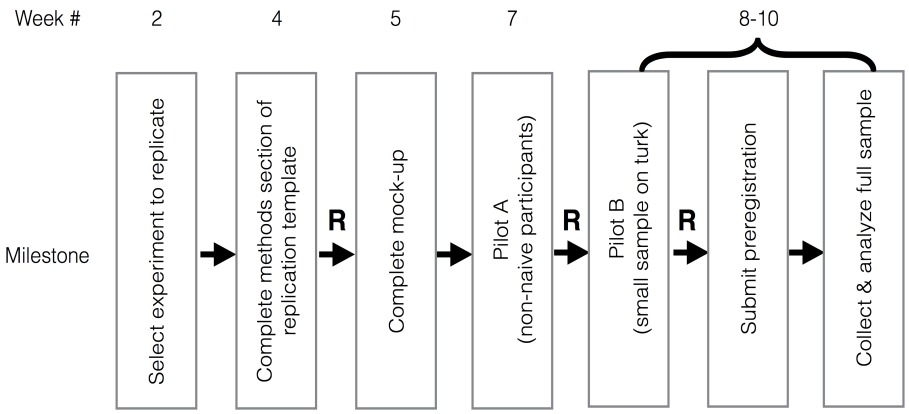

Figure 1. A schematic view of our class timeline for replication projects. R indicates the approximate timing of instructor team reviews of student materials.

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 transitions between locations on smartphone apps (Liverence & Scholl, 2015). A second cluster came from studies of social perception and social 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 Supplemental Material.

Participants

All study participants were recruited on Amazon Mechanical Turk (AMT). Individual sample sizes are given in Table 1. Each sample was recruited independently, using the same account but a different title. Of the 11 studies included in the final sample, 5 (45%) were originally conducted on AMT. For the other studies in our sample, demographic differences (in terms of age, sex, socio-economic status, and in some cases, national origin) are an important factor in interpreting our findings; see Supplemental Material for more discussion of this issue in individual cases.

We determined sample sizes using a case-by-case selection criterion that attempted to maximize the success of the project while staying within the constraints of our budget (see below for discussion).[1] For all criteria, experimenters recruited an additional 5% of participants to ensure we still achieved the desired criterion if some participants skipped through or did not complete the study. In four cases, we powered the replication attempt to 80% power based on post-hoc power calculations of the original effect sizes. In four cases, we used the original sample size. In one case, the original study was powered well above what would be required to find the effect of interest based on the post-hoc analysis, so we chose a smaller sample size that yielded sufficiently high power for that effect. In two others, we reduced the number of trials as part of the adaptation to Mechanical Turk (in one of these, changing the sample size to compensate). In the last case, the original study used a multi-step procedure where the first step consisted of stimulus generation; we decreased the N in the first stage of the procedure and focused on achieving high power at the second stage.

[1] 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.

Materials and Methods

Students used JavaScript and HTML/CSS to reimplement their respective target studies in the web browser, following the methods specified by the original authors as closely as possible. Of the 11 studies included in the final sample, 64% of studies had either some or all materials openly available. For 6 studies, 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 (this template is included in our course materials). Responses were received in all but one case, all within a matter of days.

Milestones, Preregistration, and Review

In addition to as-needed guidance on projects, students advanced through a formal review process in which each student's 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, students wrote a brief proposal selecting their replication targets, which were reviewed for feasibility and concreteness. After their selection was approved, students proceeded to write the methods section of the Open Science Collaboration (2015) replication report template. This section included a power analysis, proposed sample size, and explicit description of differences from the original study (see course materials for complete template). At the same time, students worked on an experiment 'mock-up': a fully functional but unpolished outline of their experiment, potentially using placeholders for one or more unfinished elements of the design. Upon reviewing the completed methods and mockup, the teaching staff gave thorough feedback and suggested changes to be made before beginning any data collection.

Further reviews were coordinated with completion of two pilot samples. 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 accurately logged and that analytic scripts for the confirmatory analyses functioned appropriately. After Pilot A was completed, the instructor or a TA critiqued and reviewed the student's experimental script, analytic code, and resulting output.

Once requested changes were made, the student conducted "Pilot B," using a handful of naive participants recruited from AMT. The goal of this pilot was to ensure that participants were able to complete the study and did not report any substantial issues with instructions or technical details of the study (all 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 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 for the key test selected in the original article, and that all relevant exclusion criteria, manipulation checks, etc. specified by the original authors were implemented correctly.

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.

Statistical Approach

Despite the general emphasis in meta-analytic and reproducibility efforts on aggregating a single effect size of interest (Lipsey & Wilson, 2001; Open Science Collaboration, 2015), it was often challenging for us to identify a single key statistical test for each study. Usually a substantial number of statistical tests were conducted,

often within one or several regression models with multiple specifications. Thus, we interpret "key statistical test" results with caution (see Empirical Results and Discussion below). Despite this interpretive difficulty, we followed previous work and attempted to use a single statistical test as our target for replication planning and analysis.

Pedagogy: Results and Discussion

Results: Pedagogical Assessment

All of the 15 students in the class completed projects (two with extensions going beyond the class period). Of these, one chose a project outside of the sampling frame, and three opted out of the broader project prior to data collection.[2] The remaining 11 students contributed code, data, registrations, and materials to the final product. (In one case, a student did not complete the pre-registration procedure correctly but still submitted a pre-specified 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: 1) limited time for iterative piloting and adjustment, 2) limits on funding for larger samples, and 3) limits on 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 replications with generally good statistical power. Thus, our work here shows what is possible for motivated students using freely-available tools and resources.

[2]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.

Discussion: 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 entry and encourage more researchers and instructor to include replications 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 others' courses that incorporate replication projects will look quite different from our own, we recommend having the replication as a centerpiece of the course and giving ample opportunity for feedback throughout the study design. Without a central focus on the replication project, students will not have the time or motivation to complete a high quality project; without multiple opportunities for extensive (and often interactive) instructor feedback, the quality of the final projects 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 different possible models of replication projects, for general undergraduate classes, advanced undergraduate/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. The sections below provide some discussion of individual choice-points.

	Classroom Models		
Course Attributes	Model 1	Model 2	Model 3
Student group	General Undergraduate	Advanced Undergraduate or Graduate	Sophisticated Graduate
Learning goal	Gain experience with research process	Gain research independence	Master best practices and tools
Group size	Full class / medium-size groups (7-8 students)	Small groups (2-4) or individuals	Individual
Project selection	Single class replication or small curated list	Curated list	Open project selection
Subject population	University or classroom participant pool	In-person or online convenience sample	Targeted sampling of original populations
Workflow tools	Use GUI tools / instructors implement studies	Teach one key open-source tool (e.g., R)	Teach full ecosystem of key open-source tools
Pre-registration	Pre-registration by instructor	Pre-registration by instructor (drafted by students)	Pre-registration by students (with instructor review)
Dissemination	If meeting pre-specified quality standard	With instructor approval	As class default (subject to discussion)

Table 2

Three potential models for incorporating replication projects into classes of different levels.

Project selection. In our class, students selected projects based on interest, rather than via a systematic sampling strategy. After selecting a paper, students chose a particular study from within that paper based on their judgment of interest and feasibility. We encouraged students with little programming background to opt for methodologically simple studies, and more advanced students to take on a challenging design. 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 replication projects that students can select from. This step would both give instructors a greater degree of familiarity with the literature and allow for the development of more scaffolding with respect to the development of experimental materials. A mid-level course might split the difference between these two models, providing students with a broader – but still curated – list of potential projects.

Subject population. We chose to use Amazon Mechanical Turk (AMT) as a platform for our replication studies.[3] We used AMT to facilitate the process of recruiting samples large enough to enable replications of between-subjects designs (which typically require large samples) as well as to help students learn about a valuable resource for recruitment whose use requires some specialized knowledge. As much of psychological

[3]Repeat administration of empirical work on AMT has been raised as an issue in some prior work (e.g., Chandler, Mueller, & Paolacci, 2014), but 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 failures to distinguish between related experimental paradigms. In addition, half of the AMT population is estimated to change every six months (Stewart et al., 2015), and we expected that most of the studies in our sample that had been conducted on AMT had been performed at least a year previously (though see Supplemental Material for discussion of one particular paradigm that has been used more recently).

research is now conducted online, an additional pedagogical benefit of conducting online replications is to build competence in these skills. Even if students' current research does not use AMT, these replications allow students to build a skillset that can be used throughout their academic careers. Our general class structure does not depend on the use of AMT, however, and the use of this particular online sample should be viewed as limiting the interpretation of some of our results. Other classes many find participant pools or recruitment of community samples more feasible and appropriate.

Workflow tools. All our studies 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. 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: each milestone in our course, corresponding to each phase of a research project, required students to fill in additional sections of single unified document. 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 document.[4].

Some of these tools can feel inaccessible or intimidating to students with a more limited programming background. Indeed, our students, who came from a variety of disciplinary backgrounds, typically have relatively little experience with web-based empirical studies or literate, reproducible data analysis. Yet by the end of the course, all

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

gain sufficient proficiency with the suite of technical and conceptual tools necessary 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 that students can learn to use a broad range of open 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 research methods course, instructors must gauge whether the added difficulty of a programming component is appropriate for the skills and backgrounds of their students.

Sample size and power analyses. Determining appropriate sample sizes is a major challenge in replication research (Button et al., 2013; U. Simonsohn, 2015). Sample planning based on analysis of the achieved effect size in a previously-published study is problematic because of the likelihood of inflation of published 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, U. 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 studies that were initially large and/or adequately powered. In general, samples for course projects will be limited by the available resources. For all replication research of the type advocated here, achieving high power for a smaller number of studies, perhaps conducted by

groups, is preferable to conducting more lower-power studies, because results are much more likely to make a contribution to the replication literature (as well as to be interpretable by students). Regardless of the eventual sample collected, we find that the discussion of sample size determination is one of the most eye-opening parts of our class and so we encourage instructors to allow students to participate in this planning process.

Author contact. Over half of our students contacted original authors to request study materials or clarification on methods and/or analysis, but for the purposes of our class, we chose not to require contact with all original authors. We strongly believe methods and materials should be open to allow for replication and productive science (Nosek et al., 2015), and hoped to empower students to engage with and build on the published literature directly – without personal relationships or even personal contact. That is, we chose to do good-faith replications given all information in the original articles and supplemental online materials. Authors were contacted only in cases where there were issues about access to materials or ambiguities in design or analysis. There are benefits of contacting original authors, however. In addition to professional courtesy, authors can provide valuable feedback or guidance (albeit at the cost of some imposition) prior to data collection, potentially heading off 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 in this area.

Ethical approvals. Ethical review standards vary across countries and are idiosyncratic across institutions, but we suspect certain ethical concerns will commonly arise when attempting to implement class replication projects. First, instructors may first be concerned that ethics approval is not possible within the time constraints of the class. To mitigate this issue, we suggest contacting your ethics body well in advance of the class to determine whether to create a standard protocol that encompasses all

replications that will be conducted (as we did), or to determine a timeline for submitting individual protocols very early in the course. Both types of protocol will be simplified if the study or pool of studies to be replicated is pre-selected by instructors. Second, review boards may be concerned about risks particular to the collection of data by students, e.g., lack of debriefing or greater risk of breach of confidentiality due to novice experimenters. We recommend building training around ethical issues into the syllabus of replication-based classes. This practice is both positive for students and mitigates potential ethics approval concerns via the inclusion of explicit training.

Empirical Findings: Results and Discussion

Results

All reported results are from pre-registered, confirmatory analyses.[5]

Subjective fidelity of replications. Each member of the instructor team coded the fidelity of replications, weighing whether the materials, sample, and task parameters differed from the original studies. We coded projects independently on a scale of 1 – 7 (with 1 being a loose replication with substantive deviations from the original, and 7 being essentially identical), and then discussed our ratings and made adjustments (without coming to full consensus). The mean rating was 5.29 (range: 4 –

[5]In two cases, in final review of the projects (after data analysis), 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, 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 of interest. We report results from these corrected tests here. We return to this issue below, as we believe that test selection is a critical theoretical issue in replication research (see e.g., Monin (2016) for discussion).

6.67). Several studies were essentially identical to the originals, but there was a group of others that included differences in population, number of trials, or method that might plausibly have had an effect on the results.

Subjective success of replications. After data analysis, for each replication, the student carrying out each study and the instructors all gave independent "replication ratings" to assess the subjective success of the replication. Our three-point rating system was based on the theoretical support for the original finding, "None" (no support; 0%), "Partial" (some support; 50%), and "Full" (replication consistent with original interpretation; 100%). 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 55% and 50% respectively.

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. 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 slightly larger probability of being judged successful if they were judged to be closer to the original (above the median subjective rating; 70% vs. 42%). There were also small numerical effects of the original study 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 (36%) 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 conclusive, 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 supplement inferences drawn from p-values in different ways. The effect size is an estimate of the *magnitude* of an effect. Unlike the p-value, this underlying magnitude is stable across sample sizes (assuming a constant experimental procedure) though the estimate may be more or less noisy. It thus allows for a more fine-grained comparison across studies than a binary significance criterion. The Bayes Factor is an alternative hypothesis testing method that directly quantifies the evidence in favor of one hypothesis relative to another (Jeffreys, 1961). Among other advantages, it does not privilege the null hypothesis and can provide evidence either for or against a hypothesis (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 additional criteria for evaluating replications: If the replication shows roughly the same effect size or Bayes factor as the original results despite p-values on different sides of the arbitrary $p = 0.05$ threshold, we may nonetheless consider it more of a success. 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.[6] 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

[6]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 formulae are given in Open Science Collaboration (2015), Supplemental Information.

smaller in the replications (60% of the original; 95% CI [37 - 84]), consistent with findings from multi-study replication projects (e.g. Open Science Collaboration, 2015).

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 the lack of consensus on how to conduct such a test (Anderson et al., 2016; Gilbert, King, Pettigrew, & Wilson, 2016), recent concerns that using confidence intervals in this manner may lead to misleading interpretations (Morey, HoekstraLee, & Wagenmakers, 2016), and discrepancies in the authors' 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 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.[7] This analysis revealed a number of interesting findings (Fig. 2B). 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. Finally, the Bayes Factors for several of the original effects did not appear to show strong evidential value (e.g., BF < 3, indicating that the alternative hypothesis is less than three times more likely than the null).

Discussion: Implications for Replication Research

Statistical Findings. From a purely statistical point of view, the results of our studies were underwhelming. Despite our relatively close adherence to the original

[7] This default test assumes an alternative hypothesis that the effect is present with effect size $d = .707$; a default BF for one paper could not be computed based on the statistical test that was used.

protocols and relatively large sample sizes, the modal outcome of our projects was partial replication. These partial replications often were cases 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. We invite readers to browse the narrative descriptions of individual replication attempts in our Supplemental Material to see the ways that patterns of empirical findings can differ from one another beyond the significance of a single test.

However, our statistical findings mask a more optimistic message: In most of our projects, the next step for a motivated experimenter is clear. For example, in some projects we suspect that a followup could find strong evidence for the phenomenon of interest by titrating the particular planned analysis or the difficulty of the stimulus materials. In others, the difficulty appeared to be statistical power, since differences were in the predicted direction and sometimes reached significance in subsidiary analyses, so a followup would likely require a larger sample. And in some, differences of population would mean that followups would require further stimulus or task adaptation. 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, in press for an example of this strategy and some discussion).

Many statistical factors leading to lowered replicability have been discussed in past work, including 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 both reasons of design and scale, our study here cannot disentangle these factors empirically. Nevertheless, the decreases in effect size and evidential value we report seem consistent with two explanations. First, we likely saw a "winner's curse," with initial publications tending to over-estimate

effects. Second, we also might have seen a reduction in effect size because the original studies were tailored to their specific context and population, whereas our replications may not have been (Van Bavel et al., 2016, but cf. Inbar (2016)). Future work would benefit from 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 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 key statistical test – and its associated effect size – being the properties of a study that can be targeted for replication. Yet, as we mentioned above, selecting this "key 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, no study we attempted to replicate was pre-registered (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. Almost every study conducted multiple statistical tests, often associated with several tersely-reported statistical models with differing specifications. And in some 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, including our own – makes it extremely challenging to do replication research focused on the estimation of a single

effect size.

General Discussion

In this paper, we reported a series of 11 student replications of previously-published empirical studies from 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. Importantly, rather than attempting to gauge the truth of particular effects, our project 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).

This kind of cross-class collaboration would require instructors to make class results more broadly discoverable. We can envision a number of possible avenues for this kind of sharing. For example, if some products are shared openly in a repository like the Open Science Framework (osf.io) or figshare (figshare.com) and tagged appropriately, they will be discoverable via search engines. Alternatively, a more structured method for sharing and discovery would be upload to specific replication curation websites like PsychFileDrawer (psychfiledrawer.org) or CurateScience (curatescience.org). Both of these models assume limited coordination across instructors, but more structured collaborations are of course possible. For example, the Collaborative Replications and Education Project (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.

One challenge for this kind of dissemination is how to curate which projects are released publicly. We are generally optimistic about the contributions of student work,

but of course not every student project will be scientifically informative. To take an edge case, imagine a student makes an error in stimulus presentation that is not caught by the instructors until after data collection. In this case, we would argue that this work should not be deposited in a targeted replication repository. But in the end, these decisions will be the responsibility of the course instructor – much like decisions about when to publish other research.

In sum, 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 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.

Supplemental Material: Project-By-Project Methods Addendum

For each project, we present a summary of the finding, the key statistical test, the impressionistic outcomes of the replication, and – where applicable – our assessment of the source of differences in results (often composed with the advice of the original authors). Figure 3 shows a side-by-side comparison of the key visualization for each study.

Atir, Rosenzweig, & Dunning (2015)

This paper investigated how perceived knowledge affects "overclaiming" (claiming knowledge that you do not have). We replicated 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 reported 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).

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), see below – because materials were collected in Experiment 1 (and were available for reuse), 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 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.

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. We replicated 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. The original study reported this analysis both with and without demographic controls. We chose as our key analysis the version of this regression which controlled for age, education, and income (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 when including demographic controls. However, when we did not include demographic controls, we found a marginally-significant relationship. In exploratory analyses, we also observed an unpredicted effect of income such that participants who reported higher income selected to start saving sooner.

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. Experiment 1, our target for replication, 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 original authors did not report training effects either within or across testing epochs (groups of 50 trials), so in our adaptation we decreased the length of the paradigm from four to two epochs.

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 controlling for learning across the duration of the experiment. We failed to find this overall difference. In an exploratory followup analysis using a linear mixed effects model, however, we did find a highly-significant effect of condition (when controlling for learning epoch). Our data may have been more variable for a number of reasons: unlike in the original paradigm, our participants were not required to click when they found targets; our stimuli differed slightly (the images were different and featured thin black borders); and our participants differed in age and educational status from the undergraduates in 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, our replication target, 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 of our replication selected a secondary analysis as the key test statistic (an ANOVA over two of the four cells), 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.

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 replication 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 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. 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.[8] 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

[8]We note that this project was submitted late and the student failed to preregister this analysis formally; the pre-written analysis protocol went through pre-data collection instructor review just as with other projects, however.

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. Consistent with the original report, though, we found a main effect such that stage one participants predicted that raters would like them more than the stage two participants actually did. We may have suffered from low power due to the relatively small number of profiles we collected in the replication study. In addition, a small number of our profile writers produced language in a manipulation check suggesting that the "maximize interest" manipulation had not succeeded fully; since there was no specified exclusion criterion in the original study, we included these participants.

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 replication 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. The original sample was Israeli women, and typical faces were constructed to be typical for that sample. Rather than attempting to construct a typical face for AMT workers, we instead used the stimuli provided by the original authors, recognizing that this decision limited the interpretation of our findings.

While the original course project 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 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. 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. This result was plausibly caused by the fact that the "typical" faces were probably not typical for many of our AMT participants (who were of both genders and likely from a diverse range of ethnic backgrounds).

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. There was a main effect of save condition, however, hence the authors' key manipulation succeeded for both load conditions. Perhaps even two words was sufficient to create interference for our online sample.

Wang et al. (2015)

This paper reported that the relationship between math anxiety and math performance follows an inverted-U pattern among students with higher intrinsic math motivation, whereas this same relationship is linearly negative among students with

lower intrinsic math motivation. Our replication target was Study 2, the authors' own replication study, in which college undergraduates 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. Population differences provide one plausible explanation for the differences we found (especially since the original study was itself a replication): Math performance was overall lower in our AMT sample than in the original undergraduate sample. It is possible that either the inverted U was undetectable given this lower level of performance, or that levels of motivation differed enough that the posited relationship did not hold.

Xu & Franconeri (2015)

This paper explored constraints on visual working memory in the context of mental rotation. In Experiment 1a, our replication target, 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 AMT participant population led to this floor effect.

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. We enforced a four minute writing time and a 20 word minimum; we found that most participants stayed relatively closer to these limits than participants in the original study, leading to an overall shorter priming phase with less writing.

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.

Chandler, J., Mueller, P., & Paolacci, G. (2014). NonnaÃŕvetÃĭ among amazon mechanical turk workers: Consequences and solutions for behavioral researchers. *Behavior Research Methods*, *46*(1), 112–130.

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., & 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.

Inbar, Y. (2016). Association between contextual dependence and replicability in psychology may be spurious. *Proceedings of the National Academy of Sciences*, 8676.

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.

Koole, S. L., & Lakens, D. (2012). Rewarding replications: A sure and simple way to improve psychological science. *Perspectives on Psychological Science*, *7*(6), 608–614.

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.

Makel, M. C., Plucker, J. A., & Hegarty, B. (2012). Replications in psychology research: How often do they really occur? *Perspectives on Psychological Science*, *7*(6), 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., Jeffrey N, Lee, M. D., & Wagenmakers, E.-J. (2016). The fallacy of placing confidence in confidence intervals. *Psychonomic Bulletin & Review*, *23*(1), 103–123.

Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S., Breckler, S., ... others. (2015). Promoting an open research culture. *Science*, *348*(6242), 1422–1425.

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.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant. *Psychological Science*. http://doi.org/10.1177/0956797611417632

Simons, D. J., Shoda, Y., & Lindsay, D. S. (2016). Constraints on generality (cog): A proposed addition to all empirical papers.

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.

Stewart, N., Ungemach, C., Harris, A. J., 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*(5), 479.

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.

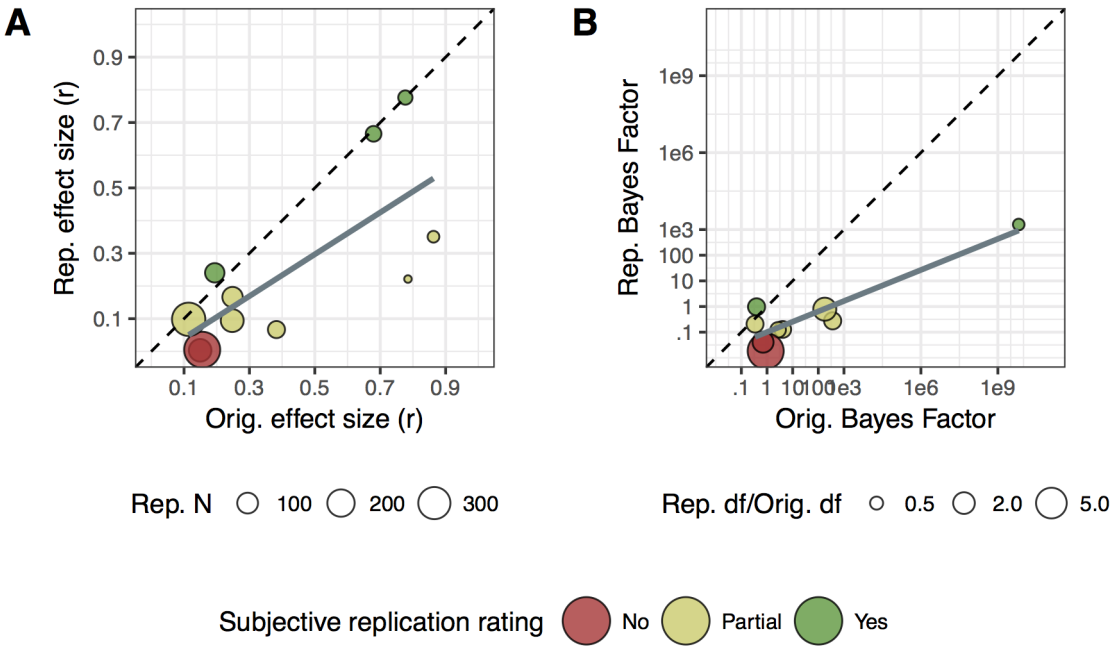

Figure 2. Replication effect size (A) and Bayes factor (B), plotted by the original effect size and Bayes factor, respectively. Point size shows replication N (A) and ratio of test degrees of freedom (B), color indicates subjective replication assessments by the authors. Note that the key statistic for one replication was a multivariate F test, hence a comparable default Bayes factor could not be computed. Additionally, the original Bayes factor for one finding was many orders of magnitude greater than the others so it is not displayed.

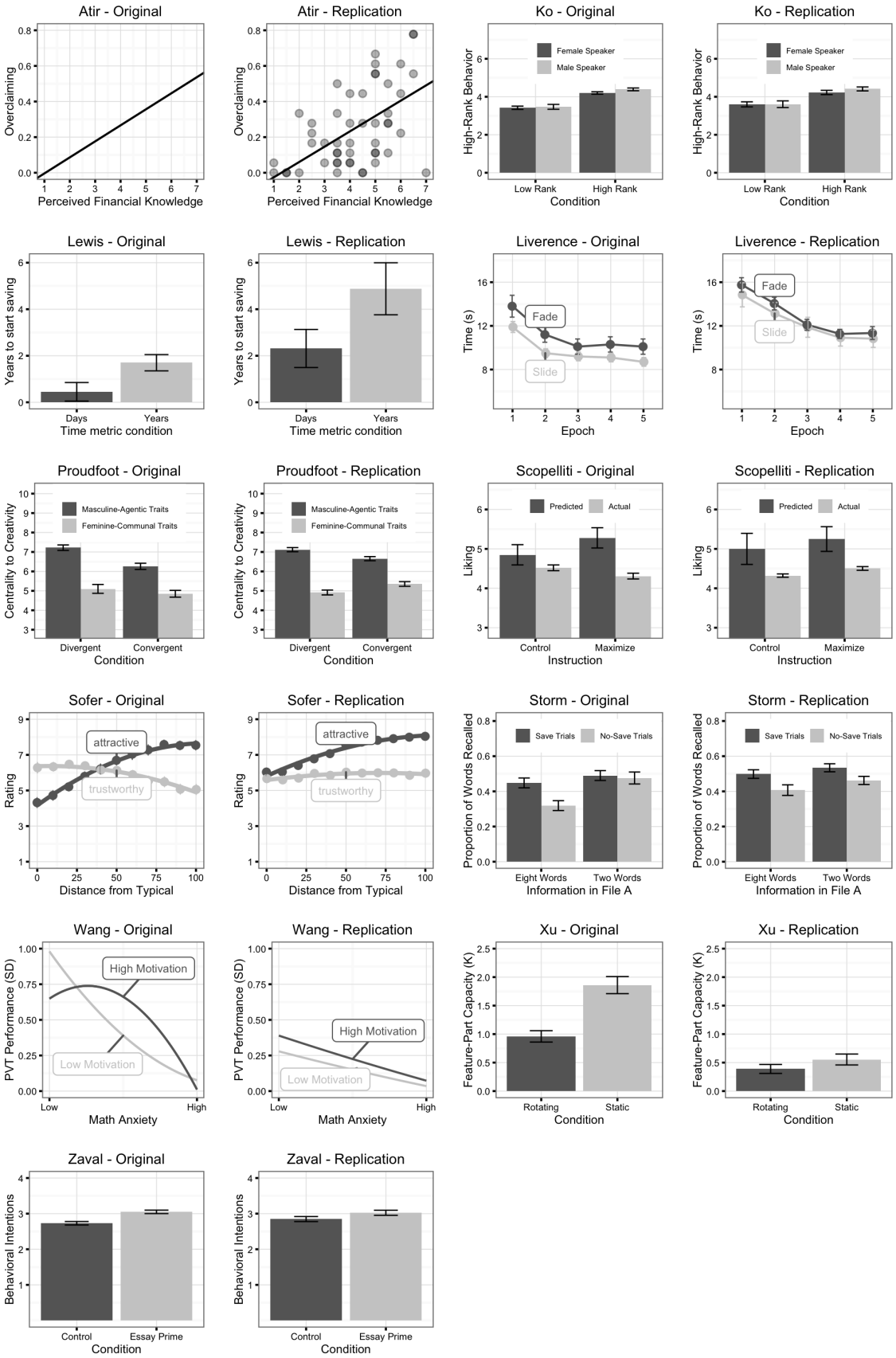

Figure 3. Side-by-side plots for each attempted replication. Error bars show standard error of the mean. Original data estimated from figures when not otherwise available.