Journal of Experimental Social Psychology 50 (2014) 217–224

Contents lists available at ScienceDirect

Journal of Experimental Social Psychology

journal homepage: www.elsevier.com/locate/jesp

The Replication Recipe: What makes for a convincing replication?

Mark J. Brandt [a,*,1], Hans IJzerman [a,1], Ap Dijksterhuis [b,2], Frank J. Farach [c,2], Jason Geller [d,2], Roger Giner-Sorolla [e,2], James A. Grange [f,2], Marco Perugini [g,2], Jeffrey R. Spies [h,2], Anna van 't Veer [a,i,2]

[a] Tilburg University, Netherlands
[b] Radboud University Nijmegen, Netherlands
[c] University of Washington, USA
[d] Iowa State University, USA
[e] University of Kent, UK
[f] Keele University, UK
[g] University of Milano-Bicocca, Italy
[h] Center for Open Science, USA
[i] TIBER (Tilburg Institute of Behavioral Economics), Netherlands

HIGHLIGHTS

- Close replications are an important part of cumulative science.
- Yet, little agreement exists about what makes a replication convincing.
- We develop a Replication Recipe to facilitate close replication attempts.
- This includes the faithful recreation of a study with high statistical power.
- We discuss evaluating replication results and limitations of replications.

ARTICLE INFO

Article history:
Received 10 July 2013
Revised 12 October 2013
Available online 23 October 2013

Keywords:
Replication
Statistical power
Research method
Pre-registration
Solid Science

ABSTRACT

Psychological scientists have recently started to reconsider the importance of close replications in building a cumulative knowledge base; however, there is no consensus about what constitutes a convincing close replication study. To facilitate convincing close replication attempts we have developed a Replication Recipe, outlining standard criteria for a convincing close replication. Our Replication Recipe can be used by researchers, teachers, and students to conduct meaningful replication studies and integrate replications into their scholarly habits.

Introduction

Replicability in research is an important component of cumulative science (Asendorpf et al., 2013; Jasny, Chin, Chong, & Vignieri, 2011; Nosek, Spies, & Motyl, 2012; Rosenthal, 1990; Schmidt, 2009), yet relatively few close replication attempts are reported in psychology (Makel, Plucker, & Hegarty, 2012). Only recently have researchers systematically reported replications online (e.g., psychfiledrawer.org, openscienceframework.org) and experimented with special issues to incorporate replications into academic publications (e.g., Nosek & Lakens, 2013; Zwaan & Zeelenberg, 2013). Moreover, some prestigious psychology journals (e.g., *Journal of Experimental Social Psychology*, *Journal of Personality and Social Psychology*, *Psychological Science*) are recently willing to publish both failed and successful replication attempts (e.g., Brandt, 2013; Chabris et al., 2012; LeBel & Campbell, in press; Matthews, 2012; Pashler, Rohrer, & Harris, in press) and even devote ongoing sections to replications (see the new section in *Perspectives on Psychological Science*, Registered replication reports, 2013).

From initial conclusions drawn from replication attempts of important findings in the empirical literature, it is clear that replication studies can be quite controversial. For example, the failure of recent attempts to replicate "social priming" effects (e.g., Doyen, Klein, Pichon, & Cleeremans, 2012; Pashler et al., in press) has prompted

* Corresponding author.
E-mail address: m.j.brandt@tilburguniversity.edu (M.J. Brandt).
[1] First two authors share first authorship.
[2] All other authors share second authorship.

0022-1031
http://dx.doi.org/10.1016/j.jesp.2013.10.005

psychologists and science journalists to raise questions about the entire phenomenon (e.g., Bartlett, 2013). Failed replications have sometimes been interpreted as 1) casting doubt on the veracity of an entire subfield (e.g., candidate gene studies for general intelligence, Chabris et al., 2012); 2) suggesting that an important component of a popular theory is potentially incorrect (e.g., the status-legitimacy hypothesis of System Justification Theory, Brandt, 2013); or 3) suggesting that a new finding is less robust than when first introduced (e.g., incidental values affecting judgments of time; Matthews, 2012). Of course, there are other valid reasons for replication failures: Chance, misinterpretation of methods, and so forth.

Nevertheless, not all replication attempts reported so far have been unsuccessful. Burger (2009) successfully replicated Milgram's famous obedience experiments (e.g., Milgram, 1963), suggesting that when well-conducted replications *are* successful they can provide us with greater confidence about the veracity of the predicted effect. Moreover, replication attempts help estimate the effect size of a particular effect and can serve as a starting point for replication–extension studies that further illuminate the psychological processes that underlie an effect and that can help to identify its boundary conditions (e.g., Lakens, 2012; Proctor & Chen, 2012). Replications are therefore essential for theoretical development through confirmation and disconfirmation of results. Yet there seems to be little agreement as to what constitutes an appropriate or convincing replication, what we should infer from replication "failures" or "successes," and what close replications mean for psychological theories (see e.g., the commentary by Dijksterhuis, 2013 and the reply by Shanks & Newell, 2013). In this paper, we provide our Replication Recipe for conducting and evaluating close replication attempts.

Close replication attempts

In general, how can one define close replication attempts? The most concrete goals are to test the assumed underlying theoretical process, assess the average effect size of an effect, and test the robustness of an effect outside of the lab of the original researchers by recreating the methods of a study as faithfully as possible. This information helps psychology build a cumulative knowledge base. This not only aids the construction of new, but also the refinement of old, psychological theories. In the definition of our Replication Recipe, close replications refer to those replications that are based on methods and procedures *as close as possible* to the original study. We use the term *close replications* because it highlights that no replications in psychology can be absolutely "direct" or "exact" recreations of the original study (for the basis of this claim see Rosenthal, 1991; Tsang & Kwan, 1999). By definition then, close replication studies aim to recreate a study as closely as possible, so that ideally the only differences between the two are the inevitable ones (e.g., different participants; for more on the benefits of close replications see e.g., Schmidt, 2009; Tsang & Kwan, 1999).

The Replication Recipe

What constitutes a convincing close replication attempt, and how does one evaluate such an attempt? This is what the Replication Recipe seeks to address. The Replication Recipe is informed by the goals of a close replication attempt: Accurately replicating methods and estimating effect sizes and evaluating the robustness of the effect outside the lab of origin. Our discussion is based on a synthesis of our own trials and errors in conducting replications and guidelines recently developed for special issues and sections of psychology journals (Nosek & Lakens, 2013; Open Science Collaboration, 2012; Registered replication reports, 2013; Zwaan & Zeelenberg, 2013). In this synthesis, we make explicit the expectations and necessary qualities of a convincing replication that can be used by researchers, teachers, and students when designing and carrying out replication studies.

A convincing close replication *par excellence* is executed rigorously by independent researchers or labs and includes the following five additional ingredients:

1. Carefully defining the effects and methods that the researcher intends to replicate;
2. Following as exactly as possible the methods of the original study (including participant recruitment, instructions, stimuli, measures, procedures, and analyses);
3. Having high statistical power;
4. Making complete details about the replication available, so that interested experts can fully evaluate the replication attempt (or attempt another replication themselves);
5. Evaluating replication results, and comparing them critically to the results of the original study.

Each of these criteria is described and justified below. We present and explain 36 questions that need to be addressed in a solid replication (see Table 1[3]). This list of questions can be used as a checklist to guide the planning and communication of a study and will help readers and reviewers to evaluate the replication, by understanding the decisions that a replicator has made when designing, conducting, and reporting their replication. These questions are intended to help replicators follow the Replication Recipe and determine when and why they have deviated from the five Replication Recipe ingredients.

Ingredient #1: Carefully defining the effects and methods that the researcher intends to replicate

Prior to conducting a replication study, researchers need to carefully consider the precise effect they intend to replicate (Questions 1–9), including the size of the original effect (Question 3), the effect size's confidence intervals (Question 4) and the methods used to uncover it (Questions 5–9). Although this can be a straightforward task, in many studies the effect of interest may be a specific aspect of a more complicated set of results. For example, in a 2×2 design where the original effect was a complete cross-over interaction, such that an effect was positive in one condition and negative in the other, the effect of interest may be the interaction, the positive and negative simple effects, or perhaps just one of the simple effects. On other occasions, the information about the methods used to obtain the effect will be unclear (e.g., the original country the study was completed in, Question 7); in these cases, it may be necessary to ask the original authors to provide the missing information or to make an informed guess. It is important to know the precise effect of interest from the beginning of the design-phase of the replication because it determines nearly all of the decisions that follow. A related consideration, especially when resources are limited, is the importance and necessity of replicating a particular effect (Question 2). Such decisions to replicate or not should be based on either the effect's theoretical importance to a particular field or its direct or indirect value to society. Another consideration is existing confidence in the reliability of the effect; an effect with a number of existing close replications in the literature may be less urgent to replicate than one without any such support (see discussion of the Replication value project, 2012–2013). In other words, not every study is worth replicating. By considering the theoretical and practical importance of a finding the best allocation of resources can be made.

Ingredient #2: Following exactly the methods of the original study

Once a study has been chosen for replication, and the precise effect of interest has been identified, the design of the replication study can commence. In an ideal world, the methods of the original study

[3] Also available as a pre-registration form on openscience.org

Table 1
A 36-question guide to the Replication Recipe.

The Nature of the Effect	
1. Verbal description of the effect I am trying to replicate:	
2. It is important to replicate this effect because:	
3. The effect size of the effect I am trying to replicate is:	
4. The confidence interval of the original effect is:	
5. The sample size of the original effect is:	
6. Where was the original study conducted? (e.g., lab, in the field, online)	
7. What country/region was the original study conducted in?	
8. What kind of sample did the original study use? (e.g., student, Mturk, representative)	
9. Was the original study conducted with paper-and-pencil surveys, on a computer, or something else?	
Designing the Replication Study	
10. Are the original materials for the study available from the author?	
a. If not, are the original materials for the study available elsewhere (e.g., previously published scales)?	
b. If the original materials are not available from the author or elsewhere, how were the materials created for the replication attempt?	
11. I know that assumptions (e.g., about the meaning of the stimuli) in the original study will also hold in my replication because:	
12. Location of the experimenter during data collection:	
13. Experimenter knowledge of participant experimental condition:	
14. Experimenter knowledge of overall hypotheses:	
15. My target sample size is:	
16. The rationale for my sample size is:	
Documenting Differences between the Original and Replication Study	
For each part of the study indicate whether the replication study is Exact, Close, or Conceptually Different compared to the original study. Then, justify the rating.	
17. The similarities/differences in the instructions are:	[Exact | Close | Different]
18. The similarities/differences in the measures are:	[Exact | Close | Different]
19. The similarities/differences in the stimuli are:	[Exact | Close | Different]
20. The similarities/differences in the procedure are:	[Exact | Close | Different]
21. The similarities/differences in the location (e.g., lab vs. online; alone vs. in groups) are:	[Exact | Close | Different]
22. The similarities/differences in remuneration are:	[Exact | Close | Different]
23. The similarities/differences between participant populations are:	[Exact | Close | Different]
24. What differences between the original study and your study might be expected to influence the size and/or direction of the effect?:	
25. I have taken the following steps to test whether the differences listed in #24 will influence the outcome of my replication attempt:	
Analysis and Replication Evaluation	
26. My exclusion criteria are (e.g., handling outliers, removing participants from analysis):	
27. My analysis plan is (justify differences from the original):	
28. A successful replication is defined as:	
Registering the Replication Attempt	
29. The finalized materials, procedures, analysis plan etc of the replication are registered here:	
Reporting the Replication	
30. The effect size of the replication is:	
31. The confidence interval of the replication effect size is:	
32. The replication effect size [is/is not] (circle one) significantly different from the original effect size?	
33. I judge the replication to be a(n) [success/informative failure to replicate/practical failure to replicate/inconclusive] (circle one) because:	
34. Interested experts can obtain my data and syntax here:	
35. All of the analyses were reported in the report or are available here:	
36. The limitations of my replication study are:	

(including participant recruitment, instructions, stimuli, measures, procedures, and analyses) will be followed exactly; however, our preference for the term 'close replication' reflects the fact that this ingredient is impossible to achieve perfectly, given the inevitable temporal and geographical differences in the participants available to an independent lab (for a similar point see Rosenthal, 1991; Tsang & Kwan, 1999). [4] Nonetheless, the ideal of an "exact" replication should be the starting point of all close replication attempts and deviations from an exact replication of the original study should be minimized (Questions 10–14), documented, and justified (Questions 17–25). Below we make recommendations for how to best achieve this goal and what can be done when roadblocks emerge.

To facilitate Ingredient #2 of the replication, researchers should start with contacting the original authors of the study to try and obtain the original materials (Question 10). If the original authors are not cooperative or if they are unavailable (e.g., have left academia and cannot be contacted, or if they have passed away), the necessary methods should be recreated to the best of the replicator researchers' ability, based on the methods section of the original article and under the assumption that the original authors conducted a highly rigorous study. For example, if replication authors are unable to obtain the reaction time windows or stimuli used in a lexical decision task, they should follow the methods of the original article as closely as possible and to fill in the gaps by adopting best practices from research on lexical decision tasks. In these cases, the replication researchers should then also seek the opinion of expert colleagues in the relevant area to provide feedback as to whether the replication study accurately recreates the original article's study as described.

In other cases, the original materials may not be relevant for the replication study. For example, studies about Occupy Wall Street protests, the World Series in baseball, or other historically- and culturally-bound events are not easily closely replicated in different times and places. In these cases the original materials should be modified to try and capture the same psychological situation as the original experiment (e.g., replicate the 2012 elections with the 2016 elections, or present

[4] Except, perhaps, when the data of a single experiment are randomly divided into two equal parts

British participants with a cricket rather than baseball championship). In such cases, the most valid replication attempt may actually entail changing the stimulus materials to ensure that they are functionally equivalent.[5] To ensure that the modified materials effectively capture the same constructs as the original study they can (when possible) be developed in collaboration with the original authors and the research community can be polled for their input (via e.g., professional discussion forums and e-mail lists). In some cases, depending on the severity of the change, it will be necessary to conduct a pilot study, testing the equivalence of manipulations and measures to constructs tested in the original research *prior* to the actual replication attempt. The justifications or steps taken to ensure that the assumptions about the meaning of the stimuli hold in the replication attempt should be clearly specified (Question 11).

Although there is no single conclusive replication (or original study for that matter), and no such burden should be put on an individual replication study, the replication researcher should do his or her best to minimize the differences between the replication and the original study and identify what these differences are. Questions 17–23 ask replicators to categorize which parts of the study are exactly the same as, close to, or conceptually different from the original study and to then justify the differences. All of these are imperfect categories that exist along a continuum, but this categorization task yields at least three benefits. First, reviewers, readers, and editors can judge for themselves whether or not they think that the deviation from the original study was justified. In some cases, a deviation will be clearly justified (e.g., using a different, but demographically similar, sample of participants), whereas in other cases it may be less clear-cut (e.g., replicating a non-internet computer-based lab study done in cubicles on the internet). Second, by identifying differences between replication and original studies (sample, culture, lab context, etc.) researchers and readers can identify where the replication is on the continuum from 'close' to 'conceptual.' Third, after multiple replication attempts have been recorded, these deviations can be used to determine relevant boundary conditions on a particular effect (for more elaboration on this point see Greenwald, Pratkanis, Leippe, & Baumgardner, 1986; IJzerman, Brandt, & van Wolferen, 2013).

In the process of identifying and justifying deviations from the original study, replicators should anticipate differences between the original and replication study that may influence the size and direction of the effect and test these possibilities (Question 24). For example, studies have revealed that people of varying social classes have different psychological processes related to the perception of threat, self-control, and perspective taking (among other things; e.g., Henry, 2009; Johnson, Richeson, & Finkel, 2011; Kraus, Piff, Mendoza-Denton, Rheinschmidt, & Keltner, 2012). Similarly, people process a variety of information differently when they are in a positive or negative mood (for reviews Forgas, 1995; Rusting, 1998). Conducting a replication at a university (or online) drawing students from different socioeconomic strata (SES) than the original population or in circumstances where participants tend to be in a different mood than the participants in the original study (e.g., immediately prior to mid-term exams compared to the week after exams) may affect the outcome of the replication. In this case, an individual difference measure of SES or mood could be included at the end of the replication study so as to not interfere with the close replication of the original study. Then, a statistical moderator test within the replication study's sample could help understand the degree to which differences in effects between samples can be explained by individual differences in SES or mood. This way it is possible to test if the differences identified in Question 24 impact the replication result (Question 25).

[5] To be sure, replications in this type of situation are less close than what is often meant by close replications and some people will consider these replications "conceptual replications".

Ingredient #3: Having high statistical power

It is crucial that a planned replication has sufficient statistical power, allowing a strong chance to confirm as significant the effect size from the original publication (see Simonsohn, 2013).[6] Underpowered replication attempts may incorrectly suggest original effects are false positives, impeding genuine scientific progress. Some authors have recommended that a sufficient amount of statistical power is at least .80 (Cohen, 1992) up to .95 (Open Science Collaboration, 2012). Because effect sizes in the published literature are likely to be overestimates of the true effect size (Greenwald, 1975), researchers should err conservatively, toward higher levels of power.[7]

Power calculations are one potential rationale for determining sample size in the replication attempt (Questions 10 & 11).[8] Calculating the power for a close replication study can be very straightforward for some study designs (e.g., a t-test). For other study designs, power analyses can be more complicated, and we encourage researchers to consult the appropriate literature on statistical power and sample size planning when designing replication attempts (see, e.g., Aberson, 2010; Cohen, 1992; Faul, Erdfelder, Lang, & Buchner, 2007; Maxwell, Kelley, & Rausch, 2008; Scherbaum & Ferreter, 2009; Shieh, 2009; Zhang & Wang, 2009 for useful information on power analysis). It has also been suggested that an alternative for determining sample sizes is to take 2.5 times the original sample size (Simonsohn, 2013).

Ingredient #4: Making complete details about the replication available

Close replication attempts may be seen as a thorny issue; openness in the replication process can help ameliorate this issue. As a rule, in order to evaluate close replication attempts as well as possible, complete details about the methods, analyses, and outcomes of a replication should be available to reviewers, editors, and ultimately to the readers of the resulting article. One way to achieve this is to pre-register replication attempts (Wagenmakers, Wetzels, Borsboom, van der Maas, & Kievit, 2012; for a pre-registration example see LeBel & Campbell, in press), including the methods of the replication study (Questions 10–16, 25), differences between the original and replication study (Questions 17–24), and the planned analysis and evaluation of the replication attempt (Questions 26–28). Following the completion of the replication attempt, the data, analysis syntax, and all analyses should be made available so that the replication attempt can be fully evaluated and alternative explanations for any effects can be explored (Questions 34 & 35).[9] Designing and conducting replications with as much openness as ethically possible inoculates against post hoc adjustment of replication success criteria, provides more transparency when readers evaluate the replication, gives people less reason to suspect ulterior motives of the replicator, and makes it more difficult to exercise liberty in choosing an analytic method to exploit the chances of declaring the findings in favor of (or against) the hypothesis (Simmons, Nelson, & Simonsohn, 2011; Wagenmakers et al., 2012). The information we recommend sharing, including the replication pre-registration and data, can be accomplished with the Open Science Framework (openscienceframework.org).

[6] When attempting to replicate a study that has already been the subject of several replication attempts it is desirable to base the replication power calculations and sample sizes on the average meta-analytic effect size.

[7] The high power necessary for a convincing close replication can provide a challenge for researchers that do not have access to very large samples. One option, though it does not appear to be used often, is to combine resources with other labs to collect the necessary number of participants (similar to initiatives developed by *Perspectives on Psychological Science*, Registered replication reports, 2013).

[8] Although, there are other defensible sample size justifications (see e.g., Maxwell et al., 2008)

[9] Exceptions can be made on data protection grounds (e.g., when data are difficult to anonymize, or when unable to share privileged information from companies).

Ingredient #5: Evaluating replication results and comparing them critically to the results of the original study

Replication studies are not studies in isolation and so the statistical results need to be critically compared to the results of the original study. The meaning of this comparison needs to be carefully considered in the discussion section of a replication article. It is not enough to deliver a judgment of "successful/failed replication" depending solely on whether or not the replication study yields a significant result. Replication effect size estimates (Question 30) and confidence intervals (when possible, Question 31) need to be calculated and the effect size estimate should be statistically compared to the original effect size (Question 32). Evaluating the replication should involve reporting two tests: 1) the size, direction and confidence interval of the effect, which tell us whether the replication effect is significantly different from the null; 2) an additional test of whether it is significantly different from the original effect. This helps determine whether the replication was a success (different from the null, and similar to or larger than the original and in the same direction), an informative failure to replicate (either not different from null, or in the opposite direction from the original, and significantly different from original), a practical failure to replicate (both significantly different from the null and from the original), or inconclusive (neither significantly different from null nor the original) (Question 33; for the criteria for these decisions see Simonsohn, 2013; for additional discussion about evaluating replication results see Asendorpf et al., 2013; Valentine et al., 2011). It may also be generally informative for any replication report to produce a meta-analytic aggregation of the replication study's effect with the original and with any other close replications existing in the literature.[10] It is important that a discussion of replication results and their conclusions take into account the limitations of the replication attempt and the original study and possibilities of Type I and Type II errors and random variation in the true size of the effect from study to study (e.g., Borenstein, Hedges, & Rothstein, 2007; Question 36). In evaluating the replication results, one should carefully consider the total weight of the evidence bearing on an effect.

One *testable* consideration for explaining differences in the results of a replication study and an original study are the many features of the study context that could influence the outcomes of a replication attempt. Some of these contextual variations are due to specific theoretical considerations. These may be as obvious as SES or religiosity in a sample, but may also be as basic and nonobvious as variations in room temperature (cf. IJzerman & Semin, 2009). In other cases, there may be methodological considerations, which may mean the manipulation or the measurement of the dependent variable is less accurate, such as when changing the type of computer monitor (e.g., CRT vs. LCD; Plant & Turner, 2009) or input device used (e.g., keyboard vs. response button box; Li, Liang, Kleiner, & Lu, 2010). For example, it is quite possible that the same stimulus presentation times using computer monitors of different brands or even the same brand but with different settings will be subliminal in one case, but supraliminal in another. Therefore, directly adopting the programming code used in the original study will not necessarily be enough to replicate the experience of the stimuli by the participants in the original study.[11] To be clear, these possible variations should not be used defensively as untested post-hoc justifications for why an effect failed to replicate. Rather, our suggestion is that researchers should carefully consider and test whether a specific contextual feature actually does systematically and reliably affect some specific results and whether this feature was the critical feature affecting the discrepancy in results beween the original and the replication study.

By conducting several replications of the same phenomenon in multiple labs it may be possible to identify the differences between studies that affect the effect size, and design follow-up studies to confirm their influence. Multiple replication attempts have the added bonus of more accurately estimating the effect size. The accumulation of studies helps firmly establish an effect, accurately estimate its size, and acquire knowledge about the factors that influence its presence and strength. This accumulation might take the form of multiple demonstrations of the effect in the original empirical paper, as well as in subsequent replication studies.

Implementation

The Replication Recipe can be implemented formally by completing the 36 Questions in Table 1 and using this information when pre-registering and reporting the replication attempt. To facilitate the formal use of the Replication Recipe we have integrated Table 1 into the Open Science Framework as a replication template (see Fig. 1). Researchers can choose the Replication Recipe Pre-Registration template and then complete the questions in Table 1 that should be completed when pre-registering the study. This information is then saved with the read-only time-stamped pre-registration file on the Open Science Framework and a link to this pre-registration can be included in a submitted paper. When researchers have completed the replication attempt, they can choose a Replication Recipe Post-Completion registration and then complete the remaining questions in Table 1 (see Fig. 2). Again, researchers can include a link to this information in their submitted paper. This will help standardize the registration and reporting of replication attempts across lab groups and help consolidate the information available about a replication study.

Limitations of the Replication Recipe

There are several limitations to the Replication Recipe. First, it is not always feasible to collaborate with the original author on a replication study. Much of the Recipe is easier to accomplish with the help of a cooperative original author, and we encourage these types of collaborations. However, we are aware that there are times when the replicator and the original author may have principled disagreements or it is not possible to work with the original author. When collaboration with the original author is not feasible, the replicator should design and conduct the study under the assumption that the original study was conducted in the best way possible. Therefore, while we encourage both replicators and original authors to seek a cooperative and even collaborative relationship, when this does not occur replication studies should still move forward.

Second, some readers will note that the Replication Recipe has more stringent criteria than original studies and may object that "if it was good enough for the original, it is good enough for the replication." We believe that this reasoning highlights some of the broader methodological problems in science and is not a limitation of the Replication Recipe, but rather of the modal research practices in *some* areas of research (LeBel & Peters, 2011; Murayama, Pekrun, & Fiedler, in press; Simmons et al., 2011; SPSP Task Force on Research & Publication Practices, in press). Original studies would also benefit from following many of the ingredients of the Replication Recipe. For example, just as replication studies may be affected by highly specific contexts, original results may also simply be *due* to the specific contexts in which they were originally tested. Consequently, keeping track of our precise methods will help researchers more efficiently identify the specific conditions necessary (or not) for effects to occur. A simple implication is

[10] Note that in a meta-analytic approach the overall effect size would almost certainly be affected more by a high-powered replication than the original study (assuming it had less statistical power). Under these conditions, the somewhat surprising conclusion is that one should trust the results of the *higher-powered* replication more than a lower-powered original study, assuming the replication is of high quality and there are no meaningful moderators of the differences between the original and replication study. A status quo in which most original studies reach equally high power levels would eliminate this imbalance.

[11] This example was adapted from a talk by Dominique Muller given at the 2013 European Social Cognition Network meeting.

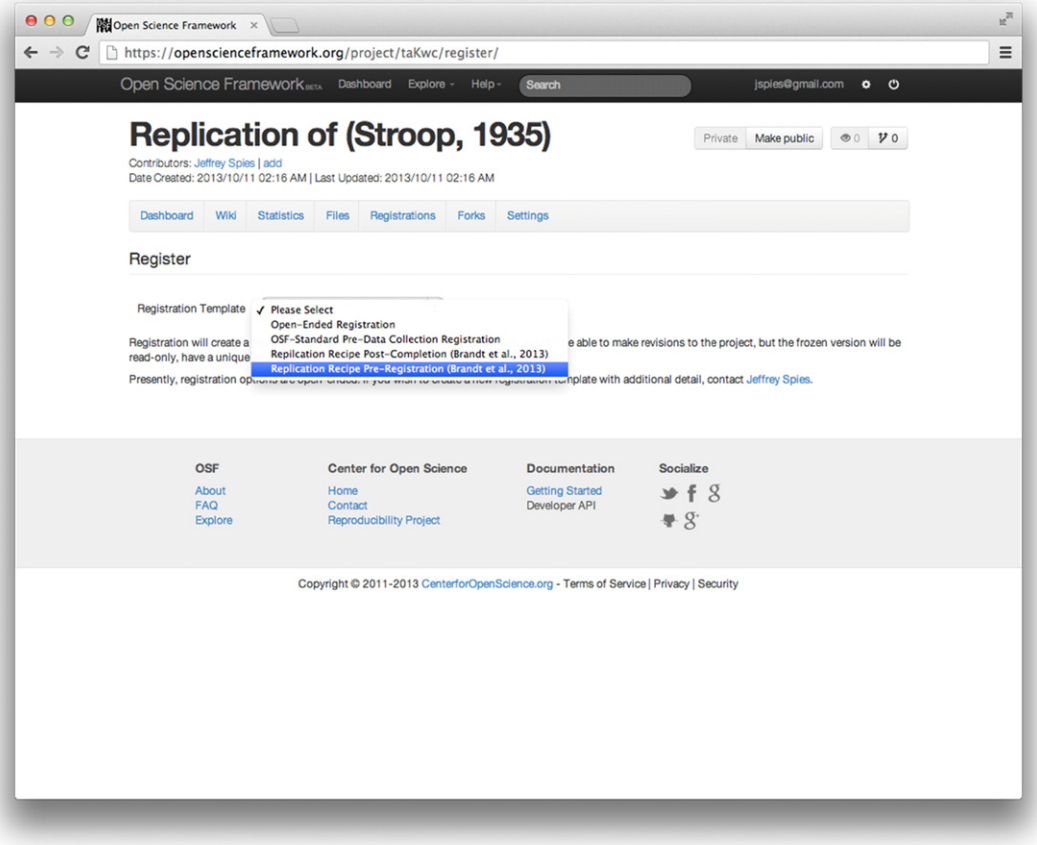

Fig. 1. Choosing the Replication Recipe as a replication template on the Open Science Framework.

that, both for replication and original studies, (more) modesty is called for in drawing conclusions from results.

Third, the very notion of single replication "attempts" may unintentionally prime people with a competitive, score-keeping mentality (e.g., 2 failures vs. 1 success) rather than taking a broader meta-analytic point-of-view on any given phenomenon. The Replication Recipe is not intended to aid score keeping in the game of science, but rather to enable replications that serve as building blocks of a cumulative science. Our intention is that the Replication Recipe helps the abstract scientific goal of "getting it right" (cf. Nosek et al., 2012) and is why we advocate conducting multiple close replications of important findings rather than relying on a single original demonstration.

Fourth, successful close replications may aid in solidifying a particular finding in the literature; however, a close replication study does not address any potential theoretical limitations or confounds in the original study design that cloud the inferences that may be drawn from it. If the original study was plagued by confounds or bad methods, then the replication study will similarly be plagued by the same limitations (Tsang & Kwan, 1999).[12] Beyond close replications, conceptual replications, or close replication and extension designs, can be used to remove confounds and extend the generalizability of a proposed psychological process (Bonett, 2012; Schmidt, 2009). When focusing on a theoretical prediction rather than effects within a given paradigm, a combination of close and conceptual replications is the best way to build confidence in a result.

Fifth, a replication failure does not necessarily mean that the original finding is incorrect or fraudulent. Science is complex, and we are working in the arena of probabilities meaning that *some unsuccessful replications are expected*. It is this very complexity that leads us to suggest that researchers keep careful track of the differences between original and replication studies, so as to identify and rigorously test factors that drive a particular effect. Indeed, just as moderators that "turn on" or "turn off" an effect are invaluable for understanding the underlying psychological processes, unsuccessful replications can also be keys to unlocking the underlying psychological processes of an effect.

Conclusion

It is clear that replications are a crucial component of cumulative science because they help establish the veracity of an effect and aid in precisely estimating its effect size. Simply stated, well-constructed replications refine our conceptions of human behavior and thought. Our Replication Recipe serves to guide researchers who are planning and conducting convincing close replications, with the answers to our 36 questions serving as a basis for the replication study. We have recommended that researchers faithfully recreate the original study; keep track of differences between the replication and original study; check the study's assumptions in new contexts; adopt high powered replication studies; pre-register replication materials and methods; and evaluate and report the results as openly as ethically possible and in accordance with the ethical guidelines of the field. We have

[12] There is some question as to whether it is appropriate to make obvious improvements to the original study, such as using a new and improved version of a scale, when conducting a close replication. We suspect that it would be better if the replication, in consultation with the original authors, used improved methods and outlined the reasoning for doing so. Running at least two replications will provide the most information: one that uses the original methodology (e.g., the old measure) and one that uses the improved methodology (e.g., the new measure). A second option is to include the change in the study as a randomized experimental factor so that participants are randomly assigned to complete the study with the original or the improved methodology. These solutions would help clarify whether the original material had caused the effect (or its absence).

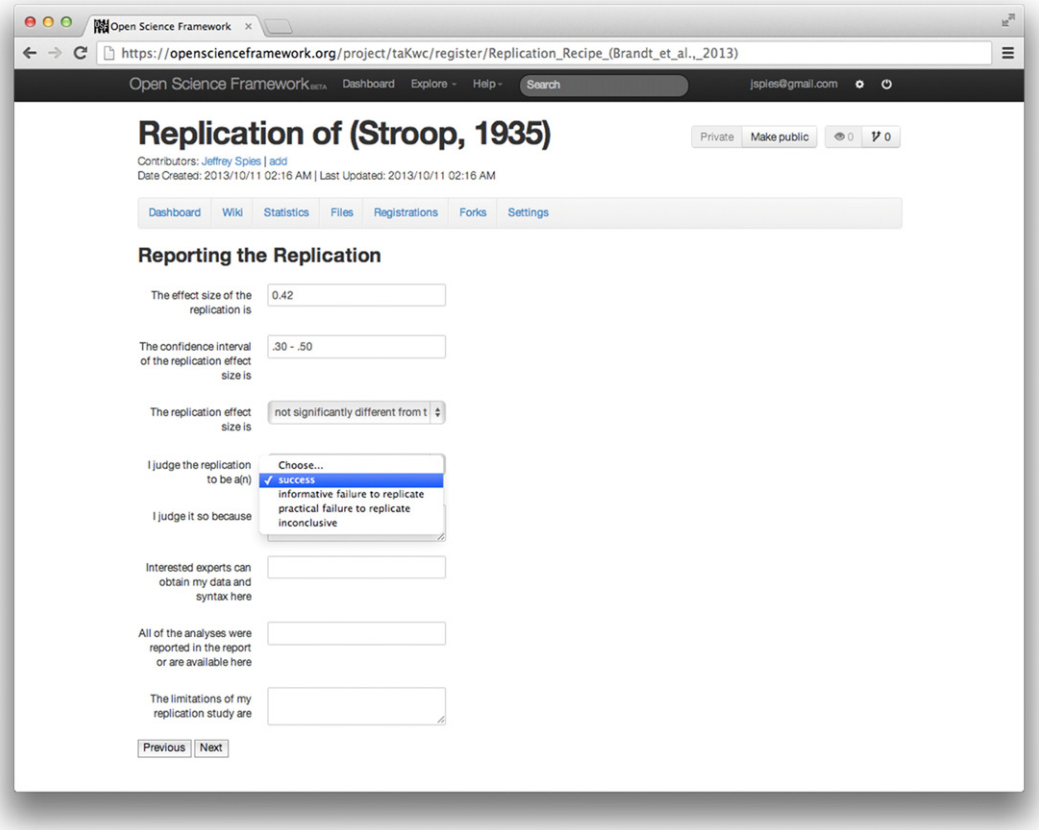

Fig. 2. Example of reporting a replication with the Replication Recipe on the Open Science Framework.

suggested that researchers measure potential moderators in a way that does not interfere with the original study, to help determine the reason for potential differences between the original and replication study, which in turn helps build theory beyond "mere" replication. By conducting high-powered replication studies of important findings we can build a cumulative science. With our Replication Recipe, we hope to encourage more researchers to conduct convincing replications that contribute to theoretical development, confirmation, and disconfirmation.

References

Aberson, C. L. (2010). *Applied power analysis for the behavioral sciences.* New York: Routledge.

Asendorpt, J. B., Conner, M., De Fruyt, F., De Houwer, J., Denissen, J. J. A., Fiedler, K., et al. (2013). Recommendations for increasing replicability in psychology. *European Journal of Personality*, *27*, 108–119.

Bartlett, T. (2013). Power of suggestion: The amazing influence of unconscious cues is among the most fascinating discoveries of our time—that is, if it's true. *The Chronicle of Higher Education* (Retrieved from http://chronicle.com/article/Power-of-Suggestion/136907/)

Bonett, D.G. (2012). Replication-extension studies. *Current Directions in Psychological Science*, *21*, 409–412.

Borenstein, M., Hedges, L., & Rothstein, H. (2007). Meta analysis: Fixed effect versus random effects. Retrieved from. http://www.Meta-Analysis.com

Brandt, M. J. (2013). Do the disadvantaged legitimize the social system? A large-scale test of the status–legitimacy hypothesis. *Journal of Personality and Social Psychology*, *104*, 765–785.

Burger, J. M. (2009). Replicating Milgram: Would people still obey today? *American Psychologist*, *64*, 1–11.

Chabris, C. F., Hebert, B.M., Benjamin, D. J., Beauchamp, J., Cesarini, D., van der Loos, M., et al. (2012). Most reported genetic associations with general intelligence are probably false positives. *Psychological Science*, *23*, 1314–1323.

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*, 155–159.

Dijksterhuis, A. (2013). Replication crisis or crisis in replication? A reinterpretation of Shanks et al. [Comment on Empirical Article e56515]. Retrieved from. http://www.plosone.org/annotation/listThread.action?root=64751

Doyen, S., Klein, O., Pichon, C. L., & Cleeremans, A. (2012). Behavioral priming: It's all in the mind, but whose mind? *PLoS ONE*, *7*, e29081.

Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*, 175–191.

Forgas, J. P. (1995). Mood and judgment: The affect infusion model (AIM). *Psychological Bulletin*, *117*, 39–66.

Greenwald, A. G. (1975). Consequences of prejudice against the null hypothesis. *Psychological Bulletin*, *82*, 1–20.

Greenwald, A. G., Pratkanis, A.R., Leippe, M. R., & Baumgardner, M. H. (1986). Under what conditions does theory obstruct research progress? *Psychological Review*, *93*(2), 216–229.

Henry, P. J. (2009). Low-status compensation: A theory for understanding the role of status in cultures of honor. *Journal of Personality and Social Psychology*, *97*, 451–466.

IJzerman, H., Brandt, M. J., & van Wolferen, J. (2013). Rejoice! In replication. *European Journal of Personality*, *127*, 128–129.

IJzerman, H., & Semin, G. R. (2009). The thermometer of social relations mapping social proximity on temperature. *Psychological Science*, *20*, 1214–1220.

Jasny, B. R., Chin, G., Chong, L., & Vignieri, S. (2011). Again, and again, and again. *Science*, *334*, 1225.

Johnson, S. E., Richeson, J. A., & Finkel, E. J. (2011). Middle class and marginal? Socioeconomic status, stigma, and self-regulation at an elite university. *Journal of Personality and Social Psychology*, *100*, 838–852.

Kraus, M. W., Piff, P. K., Mendoza-Denton, R., Rheinschmidt, M. L., & Keltner, D. (2012). Social class, solipsism, and contextualism: How the rich are different from the poor. *Psychological Review*, *119*, 546–572.

Lakens, D. (2012). Polarity correspondence in metaphor congruency effects: Structural overlap predicts categorization times for bipolar concepts presented in vertical space. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*, 726–736.

LeBel, E. P., & Campbell, L. (2013). Heightened sensitivity to temperature cues in individuals with high anxious attachment: Real or elusive phenomenon? *Psychological Science*, *24*, 2128–2130.

LeBel, E. P., & Peters, K. R. (2011). Fearing the future of empirical psychology: Bem's (2011) evidence of psi as a case study of deficiencies in modal research practice. *Review of General Psychology*, *15*, 371–379.

Li, X., Liang, Z., Kleiner, M., & Lu, Z. (2010). RTbox: A device for highly accurate response time measurements. *Behavior Research Methods*, *42*, 212–225.
Makel, M. C., Plucker, J. A., & Hegarty, B. (2012). Replications in psychology research: How often do they really occur? *Perspectives on Psychological Science*, *7*, 537–542.
Matthews, W. J. (2012). How much do incidental values affect the judgment of time? *Psychological Science*, *23*, 1432–1434.
Maxwell, S. E., Kelley, K., & Rausch, J. R. (2008). Sample size planning for statistical power and accuracy in parameter estimation. *Annual Review of Psychology*, *59*, 537–563.
Milgram, S. (1963). Behavioral study of obedience. *Journal of Abnormal and Social Psychology*, *67*, 371–378.
Murayama, K., Pekrun, R., & Fiedler, K. (in press). Research practices that can prevent an inflation of false-positive rates. *Personality and Social Psychology Review* (in press).
Nosek, B.A., & Lakens, D. (2013). Call for proposals: Special issue of social psychology on "Replications of important results in social psychology". *Social Psychology*, *44*, 59–60.
Nosek, B.A., Spies, J. R., & Motyl, M. (2012). Scientific utopia: II. Restructuring incentives and practices to promote truth over publishability. *Perspectives on Psychological Science*, *7*, 615–631.
Open Science Collaboration. (2012). An open, large-scale, collaborative effort to estimate the reproducibility of psychological science. *Perspectives on Psychological Science*, *7*, 657–660.
Pashler, H., Rohrer, D., & Harris, C. (2013). Can the goal of honesty be primed? *Journal of Experimental Social Psychology*, *49*, 959–964.
Plant, R. R., & Turner, G. (2009). Millisecond precision psychological research in a world of commodity computers: New hardware, new problems? *Behavior Research Methods*, *41*, 598–614.
Proctor, R. W., & Chen, J. (2012). Dissociating influences of key and hand separation on the Stroop color-identification effect. *Acta Psychologica*, *141*, 39–47.
Registered replication reports (2013). Retrieved June 5, 2013, from. http://www.psychologicalscience.org/index.php/replication
Replication value project [Electronic mailing list discussion] (2012-2013). Retrieved from Open Science Framework Google Group: https://groups.google.com/d/topic/opensciencefframework/Hnn33i2fSyU/discussion
Rosenthal, R. (1990). How are we doing in soft psychology? *American Psychologist*, *45*, 775–777.
Rosenthal, R. (1991). Replication in behavioral research. In J. W. Neuliep (Ed.), *Replication research in the social sciences* (pp. 1–30). Newbury Park, CA: Sage.
Rusting, C. L. (1998). Personality, mood, and cognitive processing of emotional information: Three conceptual frameworks. *Psychological Bulletin*, *124*, 165–196.
Scherbaum, C. A., & Ferreter, J. M. (2009). Estimating statistical power and required sample sizes for organizational research using multilevel modeling. *Organizational Research Methods*, *12*, 347–367.
Schmidt, S. (2009). Shall we really do it again? The powerful concept of replication is neglected in the social sciences. *Review of General Psychology*, *13*, 90–100.
Shanks, D. R., & Newell, B. R. (2013). Response to Dijksterhuis. [Comment on Empirical Article e56515]. Retrieved May 20, 2013, from. http://www.plosone.org/annotation/listThread.action?root=64795
Shieh, G. (2009). Detecting interaction effects in moderated multiple regression with continuous variables power and sample size considerations. *Organizational Research Methods*, *12*, 510–528.
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*, *22*, 1359–1366.
Simonsohn, U. (2013). Evaluating replication results. Available at SSRN: http://ssrn.com/abstract=2259879
SPSP Task Force on Research and Publication Practices (in press). Improving the dependability of research in personality and social psychology: Recommendations for research and educational practice. *Personality and Social Psychology Review* (in press).
Tsang, E. W., & Kwan, K. M. (1999). Replication and theory development in organizational science: A critical realist perspective. *Academy of Management Review*, *24*, 759–780.
Valentine, J. C., Biglan, A., Boruch, R. F., Castro, F. G., Collins, L. M., Flay, B. R., et al. (2011). Replication in prevention science. *Prevention Science*, *12*, 103–117.
Wagenmakers, E. J., Wetzels, R., Borsboom, D., van der Maas, H. L., & Kievit, R. A. (2012). An agenda for purely confirmatory research. *Perspectives on Psychological Science*, *7*, 632–638.
Zhang, Z., & Wang, L. (2009). Statistical power analysis for growth curve models using SAS. *Behavior Research Methods*, *41*, 1083–1094.
Zwaan, R. A., & Zeelenberg, R. (2013). Replication attempts of important results in the study of cognition. *Frontiers in Cognition* (Retrieved from http://www.frontiersin.org/Cognition/researchtopics/Replication_Attempts_of_Import/1461)