Teaching replication

Michael C. Frank

Department of Psychology, Stanford University

Rebecca Saxe

Department of Brain and Cognitive Sciences, MIT

Thanks to Brian Nosek and other members of the Open Science Collaboration for feedback on an earlier version. We gratefully acknowledge the students and TAs in MIT course 9.61 and Stanford course Psychology 254. This work was supported by a Packard Fellowship and an NSF CAREER award to R. Saxe.

Please address correspondence to Michael C. Frank, Department of Psychology, Stanford University, 450 Serra Mall (Jordan Hall), Stanford, CA, 94305, tel: (650) 724-4003, email: `mcfrank@stanford.edu`.

Abstract

Replication is held as the gold standard for ensuring the reliability of published scientific literature. But conducting direct replications is expensive, time-consuming, and unrewarded under current publication practices. So who will do the replications? Our answer is that students in laboratory classes should replicate recent findings as part of their training in experimental methods. In our own courses, we have found that replicating cutting-edge results is exciting and fun, it gives students the opportunity to make real scientific contributions (provided supervision is appropriate), and it provides object lessons about the scientific process, the importance of reporting standards, and the value of openness.

Introduction

Scientific progress in experimental psychology depends on accumulating a body of reliable experimental results. Reports of experiments thus ought to contain recipes for how to achieve the same results. Of course, scientists cannot afford to re-do every important experiment; most of the time, we must rely on one another's reports. Thus it is of critical importance to know whether and when we can rely on the published literature.

One approach to guaranteeing the reliability of published results is to focus on statistical standards (Cohen, 1994; Wilkinson, 1999; Sharpe, 2004). A combination of appropriate sampling procedures and statistical testing can allow researchers to estimate reliability. Either they can compute the probability that the observed data were produced via chance variation or the Bayesian reverse, the probability of a particular hypothesis given the observed data (Cohen, 1994; Wagenmakers, Wetzels, Borsboom, & Van Der Maas, 2011). But many factors are still unaccounted within either of these paradigms, notably including the number of other experiments performed by the same researchers (e.g. Rosenthal, 1979) and analytic decisions not accounted for in the statistical calculation (e.g. Simmons, Nelson, & Simonsohn, 2011).

Consequently, the gold standard for reliability is independent replication—defined here as the repetition of the experimental methods that led to a reported finding by a researcher unaffiliated with the original result. If even a small number of independent replications find the same results, then that result can be relied upon—whether in theoretical syntheses, in meta-analyses, or as the basis for designing further experiments.

Nearly everyone believes that there should be more replication of published results, and more public reporting of the results of replication attempts, but the field has come to no clear consensus about how to encourage these practices. The primary issue is the limited incentives available for replication. Successful direct replications are almost never published (Neuliep & Crandall, 1990, 1993; Madden, Easley, & Dunn, 1995). Repeated

failures to replicate are occasionally published (e.g. Ritchie, Wiseman, & French, 2012; Doyen, Klein, Pichon, & Cleeremans, 2012), but rarely in an outlet with the same impact as the original publication. A recent special issue of *The Psychologist* was devoted to this topic, but it provided limited consensus on a solution. Proposals include forums for reporting replications (Simons, 2012; Holcombe & Pashler, 2012; Hartshorne & Schachner, 2012), publication standards that encourage internal (by the same lab, rather than by an independent lab) replications (Roediger, 2012), and greater openness about data and methods for replications of analyses (Wicherts, 2012). But all of these are—at their core—requests for busy scientists to do something that is both less exciting and less rewarding than conducting new research.

In this article, we propose an alternative solution: Replications of new research should be performed by students as part of their coursework in experimental methods. In making this suggestion, we draw on our own experiences from methods courses we teach for undergraduates (at MIT) and graduate students (at Stanford). The balance of this paper describes our proposal in more depth, what we see as its merits, and some possible questions and criticisms. One goal of recent discussions of replication is to encourage a shift towards a culture in which new work in experimental psychology builds more directly on previous methods and results. Replicating and extending allows researchers to create an interlocking edifice of findings, rather than an array of unconnected phenomena (Newell, 1973). What better way to promote this kind of cultural shift than to instill our students the values that we want our young scientists to hold?

The proposal: Replications of new results in the classroom

Psychologists are motivated to conduct experiments because they want to know the answers to interesting questions about the mind. But there are many ways of getting answers of this sort. You can ask a friend or a teacher, look it up on Wikipedia, read a

book, or even consult primary scientific literature. The only reason to do an experiment—rather than getting an answer by one of these simpler and faster routes—is that there no other way to answer your question: no one knows the answer yet. Yet when we train young scientists to do experiments, we often lose track of this rationale.

Here are two caricatures of standard laboratory classed in experimental methods for psychology: In version one, students are told that a particular classic and well-established result (say, the response interference effect first documented by Stroop, 1935) is true. They are then asked to exert substantial effort acquiring the skills to verify what they already know to be true. If the students fail to reproduce the known truth, both the students and instructors infer that the students have made an error, or that uncontrolled variables, poor quality equipment, or a small sample size doomed their attempt from the beginning. In essence, students are being asked to use the experimental method to verify facts that they could have checked on Wikipedia—in the process removing the most exciting part of the experimental method: the uncertainty surrounding the possibility of a new discovery.

In version two, students are asked to choose their own question and design an original experiment to test that question. They are then given a week or two to do so. Because of the students' constraints on time and lack of expertise, the resulting experiments are often silly, poorly designed, and unlikely to connect to current issues in psychological science, and the students know it. Neither the students, nor the instructors, let alone the broader community, are genuinely curious to learn the answer to the question, and so again the excitement of discovery is absent.

We suggest that the both the instructors and the students in these courses would benefit from one small but fundamental change: instead of replicating classic experiments or creating new ones from scratch, students in laboratory classes should set out to exactly replicate recent, cutting-edge experiments. In our caricatures above, there is no motivation for care and methodological rigor. Either the answer is already known, or else finding it is

not important. This lack of motivation for the skills that are being taught may partially explain why methods courses are among the least popular course offerings, especially in the undergraduate curriculum (where they are often tied for last place with statistics).

We propose that the best way to motivate students to learn experimental methods is to use the same logic that motivates working scientists. Students have more fun, learn more, and are more careful, if they think that the answer to an interesting question hinges on getting their experiment right. But which question should they ask, and how?

Direct replications of recently published experiments have three benefits as a teaching tool. First, the original authors have already done the hard work of thinking of a novel question and designing an experiment to answer it. Second, because the paper was published recently, it is likely that the question it poses is interesting to a community of current working scientists—and thus, plausibly, to the instructor. In our experience, the instructor's genuine curiosity is infectious. Finally, the community does not yet know how reliable the results are. The students thus have an opportunity to make a real contribution to the progress of science, by answering an important question to which the answer is still unknown: if this experiment were conducted again in a new sample by an independent researcher, would the results replicate?

In our proposal for a replication-based laboratory class, students and their instructor work together to identify recently published experiments that are of interest to both parties and are feasible to replicate, given the students' skills and access to resources. This process can involve the course instructor creating a "menu" of options (generally the best case for an undergraduate class), or can involve students finding results that are relevant to their own interest (perhaps better for a graduate class). Next, students create a proposal for an direct replication, including analyses of power and discussion of any inevitable differences between the original and replication experiments that might undermine the probability of replication. When possible, the students should use the same

materials as the original authors.

The students and instructor then work together to conduct the experiment. During lab time, data are collected from a university credit pool, or from Amazon Mechanical Turk (Buhrmester, Kwang, & Gosling, 2011), or from paid participants if funds are available. The instructor and students collaborate to analyze the data and interpret whether they have replicated the finding. Based on the judgment of the instructor, at this point the results may be submitted to a replication registry (Holcombe & Pashler, 2012; Hartshorne & Schachner, 2012; Spellman, 2012). Finally, the culmination of the project for the students is a report on this replication, centered around a real scientific inference: have the students replicated the findings of the original authors? If the project was well designed and well conducted, this uncertainty makes the experiment a valuable contribution to our knowledge, and makes the student into a scientist.

Merits of student-led replications

Both the students and the broader scientific community have a lot to gain from in-class replications. Most obviously, the scientific community would get a large captive workforce. There are many more students of psychology around the world than there are working scientists. These students are already spending their time in experimental methods courses; our proposal would put their time and effort to scientific benefit. Moreover, the in-class setting provides project grades as a simple and cost-free external incentive for good work, if internal incentives such as curiosity are not motivating enough.

The benefits to the students may seem less tangible, but we suggest they are more profound. Both of us have taught project-based methods classes that are centered around replications of recent findings. These classes are challenging: they must be as small as a seminar to allow for the amount of interaction between students and instructors that we have found necessary, but they take quite a bit more instructor preparation and interaction

than a typical seminar. Yet in our experience, it is incredibly motivating for students to know that they are running experiments not purely as an exercise in skill building, but for their original purpose: to find something out that couldn't be learned in any other way.

When the instructor is genuinely (i) curious about the question, and (ii) optimistic that the students' experiment could answer it, the stakes are higher. When the result might be published to the broader community, students' designs can be held to real scientific standards. When both the standards and the stakes are higher, students are more careful. Exactly like a scientist, students must reason about possible differences in methods or sample that would threaten their chances for a useful discovery. They must push themselves to recruit the necessary number of participants to complete the study. Discussions of sampling, power, counter-balancing, within-subjects designs, and other abstract methodological concerns suddenly become more concrete. These considerations are critical to the experimental method, but without true uncertainty about the result, they are empty exercises; there's no reason to get the process exactly right unless you care about the end result.

In addition to the expertise gained by conducting the experiments, students learn a lot from reading others' papers with the eye of a replicator, not just a reader. Novice readers of scientific literature can be distracted by flashy conclusions, without careful attention to the methods. Planning a replication requires focusing on the details of methods and analyses. Inferring the rationale behind the original researcher's decisions can give students insight into the process of experimental design, and scrutiny of the statistics becomes a way to anticipate how likely their own hard labor is to end in success or failure.

Along with these primary pedagogical merits of student-led replications come many additional practical and cultural benefits. We find that reading with the intent to replicate helps students appreciate a good complete scientific report. Why do we describe what version of the software was used, or how far the participant sat from the monitor?

What is the value of giving all those means and standard deviations when the discussion interpreted only the statistical difference between two groups? We find that it becomes trivial to motivate these conventions in scientific writing, when students are trying to pick out exactly what was done and what was found by a faraway researcher. Experiencing the frustrations of trying to sort out someone else's incomplete report is the best possible argument for doing a better job yourself in future. Relying on the original authors to provide materials and advice helps teach the value of open sharing.

Thus, we think the scientific community stands to gain twice from having students conduct in-class replications: first, from the results of those replications, and second, by increasing the population of those who understand and appreciate the best in scientific conventions.

Questions about in-class replications

Can students perform a "real" replication over the course of a term? Our experience is that they can. A number of factors facilitate this process. The most important challenge is to choose the right experiments to replicate. The experiments must be within the technical abilities of the students. We have found it helpful to provide a range of options, from pencil-and-paper survey studies with large samples to more complex experimental paradigms requiring fewer participants. Students who are more ambitious, more gregarious, or more technically capable can then choose projects to suite their strengths. The more complex projects are then greatly facilitated by robust scaffolding for the initial construction of the experiment. In our classes, teaching assistants (experienced graduate students) provide extensive help in creating materials and experimental paradigms.

It must also be possible to collect enough data to satisfy power analyses—otherwise a failure will be inconclusive, a frustrating outcome for all involved. We encourage careful attention to reported (or inferred) effect sizes of the original experiment, rather than

p-values. Students can work in groups large enough to support the person-hours required for data collection within the time limits of the course. For example, in our undergraduate course, which includes 6 hours of lab time/week, pairs of students working together collect data. During the three week period after the experimental paradigm is complete, they collect data from around 100 participants in short pen-and-pencil surveys or from around 20 in lab experiments. In our graduate class, we focus on web-based experiments, which makes collecting sufficient data less prohibitive in terms of time.

Finally, we rely on frequent project milestones, providing opportunity for review of the design and discussion of key details. Without frequent check-ins (and timely feedback from instructors and teaching assistants), it is easy for student projects to get derailed.

If students conduct replications, will they be valid? In some cases. Obviously, the quality of student-led replications will depend both on the students' effort and care, and on the instructor's close supervision. Supervising novice researchers is challenging, and not every project will be a success. We propose that communication of the results to the broader research community be done, in the end, by the instructors. Instructors can use their own scientific judgment to identify the projects that meet the standards of scientific research. In our experience, more than half of the projects conducted in our course are sufficiently high quality to be considered "real" replications.

Isn't performing real replications too costly? In a word, no. We find that the major cost is in terms of instructor and teaching assistant time. With the advent of free or open-source tools for data collection (e.g. PsyScope, standard web tools like HTML and JavaScript, Python, R, and others) and the use of university-licensed products (Matlab with Psychtoolbox, Qualtrics), it is possible to create a powerful, low- or zero-cost ecosystem for conducting behavioral research in the classroom. Experimental participants can also be recruited in many different ways; for example, University participant pools and tools like Mechanical Turk can be used to collect large amounts of data at relatively

low cost.

Why not perform original research in the classroom? A criticism of our proposal from an alternative point of view asks, if we are going to empower students in methods class to perform meaningful replications of recent research, why not go beyond this and ask them to create their own *original* projects? As mentioned above, the problem is that there is rarely time in a single quarter- or semester-long course. To create and run a scientifically-meaningful original experiment, students must learn a literature, devise a novel question, design a good experiment, and then complete that experiment. While these steps feel effortless to experienced scientists, even for incoming graduate students they can often extend through the first year and beyond.

Direct replications provide a jump-start: the original authors have already had the idea and designed the experiment. Students can focus on learning the skills to perform that experiment. If there is time, however, it can be very productive for students to design their own original extension projects *after* conducting the replication—for example, in a full year honors seminar. Their replication project gives students expertise in a topic, and experience with certain kinds of methods; they can then leverage this knowledge towards a more successful project of their own. (Several projects of this type from our undergraduate course have even been part of successful publications, e.g. Ichinco, Frank, & Saxe, 2009; Frank, Fedorenko, Lai, Saxe, & Gibson, 2012).

Other issues. Of course, in-class replications also face many other challenges, which are common to any proposal for increasing independent replications. For example, whether replications are conducted by working scientists or by students (or ideally, by both), we will still need a way to record the results of replications. Several recent proposals address this need with ideas for registering replications, keeping track of which articles have been replicated, and marking those results most in need of independent verification (Holcombe & Pashler, 2012; Hartshorne & Schachner, 2012; Spellman, 2012).

Student and scientist replications both could be submitted to such a system, perhaps with student replications marked as having been conducted for a class (as is already done e.g. in `psychfiledrawer.org`).

Like any scientist planning a replication, students and their supervisors will still have to make hard decisions about how close replications should be. Should the replicator struggle to use the same software, if there is a comparable package that is easier to use? Would it matter if similarly designed but distinct stimuli were used? Must the counterbalancing be the same as in the initial report? Student responses to these questions may be somewhat different than that of more experienced researchers, given the limited resources, knowledge, and time that a student will have in conducting a course project. But this is where instructors' greater experience should act as a filter. Those replications judged to be too different from the original report can be marked as such or simply not submitted to a registry.

Finally, student replications will be biased towards those studies that *can* be replicated in the course of a term. That will inevitably mean that, within the realm of behavioral research, easy survey studies will tend to be reproduced more often than complex, psychophysical designs. Physiological and neural measures will be reproduced even less often, especially due to the cost of neuroimaging and the analytic difficulties encountered along the way. Developmental research and other work with special populations will be unlikely targets. Participants will tend to be WEIRD (Henrich, Heine, Norenzayan, et al., 2010) and they will tend to be from university participant pools or other convenient sources like Amazon Mechanical Turk. But correcting these biases is a general problem for the field of psychology, and one of the best possible correctives is to train a cohort of scientists that is aware of them and aware of the value of replication.

Conclusions

Due to recent worries about the reliability of experimental psychology, there has been intense interest in encouraging replication. But the incentives for independent replication are still quite limited, and as a result, we still do not know who will conduct the replications that we all think should be happening. Our proposal here is that students studying experimental methods should conduct replications of recent results as course projects. This practice kills two birds with one stone: it creates a pedagogically rich, rewarding environment for the students to learn the basic tools of science, while at the same time ensuring the reliability of the published literature by providing large numbers of independent replicators.

In-class replications are not a panacea: There are significant obstacles facing any proposal to improve the reliability of the scientific record. But these obstacles should not prevent us from beginning somewhere. We suggest that the beginning should be in the classroom.

References

Buhrmester, M., Kwang, T., & Gosling, S. (2011). Amazon's mechanical turk a new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science*, *6*(1), 3–5.

Cohen, J. (1994). The earth is round ($p < .05$). *American psychologist*, *49*(12), 997.

Doyen, S., Klein, O., Pichon, C., & Cleeremans, A. (2012). Behavioral priming: It's all in the mind, but whose mind? *PloS ONE*, *7*(1), e29081.

Frank, M., Fedorenko, E., Lai, P., Saxe, R., & Gibson, E. (2012). Verbal interference suppresses exact numerical representation. *Cognitive Psychology*, *64*(1), 74–92.

Hartshorne, J., & Schachner, A. (2012). Tracking replicability as a method of post-publication open evaluation. *Frontiers in Computational Neuroscience*.

Henrich, J., Heine, S., Norenzayan, A., et al. (2010). The weirdest people in the world. *Behavioral and Brain Sciences*, *33*(2-3), 61–83.

Holcombe, A. O., & Pashler, H. (2012). Making it quick and easy to report replications. *The Psychologist*, *25*(5), 349.

Ichinco, D., Frank, M., & Saxe, R. (2009). Cross-situational word learning respects mutual exclusivity. In *Proceedings of the 31st annual meeting of the cognitive science society* (Vol. 31).

Madden, C., Easley, R., & Dunn, M. (1995). How journal editors view replication research. *Journal of Advertising*, 77–87.

Neuliep, J., & Crandall, R. (1990). Editorial bias against replication research. *Journal of Social Behavior & Personality*.

Neuliep, J., & Crandall, R. (1993). Reviewer bias against replication research. *Journal of Social Behavior & Personality*.

Newell, A. (1973). You can't play 20 questions with nature and win: Projective comments on the papers of this symposium. In W. G. Chase (Ed.), *Visual information*

processing. New York, NY: Academic Press.

Ritchie, S., Wiseman, R., & French, C. (2012). Failing the future: Three unsuccessful attempts to replicate Bem's retroactive facilitation of recall effect. *PloS ONE*, *7*(3), e33423.

Roediger, H. L. (2012). Twist, bend and hammer your effect. *The Psychologist*, *25*(5), 349.

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological bulletin*, *86*(3), 638.

Sharpe, D. (2004). *Beyond significance testing: Reforming data analysis methods in behavioral research.* Washington, DC: American Psychological Association.

Simmons, J., Nelson, L., & Simonsohn, U. (2011). False-positive psychology. *Psychological Science*, *22*(11), 1359–1366.

Simons, D. (2012). The need for new incentives. *The Psychologist*, *25*(5), 349.

Spellman, B. (2012). Introduction to the special section. *Perspectives on Psychological Science*, *7*(1), 58–59.

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology*, *18*, 643–662.

Wagenmakers, E., Wetzels, R., Borsboom, D., & Van Der Maas, H. (2011). Why psychologists must change the way they analyze their data: the case of psi: comment on bem (2011).

Wicherts, J. M. (2012). Share your data. *The Psychologist*, *25*(5), 349.

Wilkinson, L. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American psychologist*, *54*(8), 594.