TEACHING REPLICATION

Jon E. Grahe[2], Fiona Fidler[1], Tim Parker[3]
[1]University of Melbourne, Australia
[2]Pacific Lutheran University, Tacoma, WA, USA
[3]Whitman University, Walla Walla, WA, USA

Students (and researchers) often overestimate the information p-values provide about future replications. Over the last 5 years, large scale reproducibility projects in psychology, biomedicine and other disciplines have exposed just how wrong these interpretations can be. Some have responded with initiatives to teach replication more explicitly and further, to teach statistics through direct replication studies. I will give examples of such programs from psychology and ecology. Benefits include connecting students with real world research (rather than contrived examples), providing context for the comparison of different statistical approaches, and introducing meta-analysis and other data synthesis techniques in an integrated way. They also meaningfully contribute to the progress of science by building an open replication data bank.

INTRODUCTION

Over the last five years, large scale reproducibility projects have uncovered high rates of irreproducible results in several disciplines, most notably psychology (Open Science Collaboration 2015) and biomedicine (Freedman et al 2015), but the underlying causes exist in all fields (Ioannidis, 2005). There is resounding consensus that the causes of this crisis are *structural*, that is, that the current incentive structures of science are at odds with best practice and scientific integrity. For example, scientific publishing and research funding are both strongly biased towards novel, original, 'ground-breaking' research at the expensive of replication and self-correction. Replication studies constitute approximately 1% of the published psychological literature (Makel et al 2012) and possibly even less of the published ecology literature (Kelly 2017). Institutions reward quantity over quality by imposing selection and promotion criteria based almost exclusively on publication metrics. The widespread practice of making decisions based on the statistical threshold of $p < .05$ have left the literature in many disciplines biased and virtually bereft of statistically non-significant ('negative' or 'null') results.

In line with the causes, the proposed solutions to the crisis involve restructuring incentives and reformulating policies, both in journals and institutions. In journals, proposed solutions include changing editorial policies to encourage submission of statistically non-significant ('negative' or 'null' results) and replication studies; creating and enforcing open data policies; promoting pre-registration, and offering a Registered Report article format. Some journals have made substantial progress on these fronts, for example, *Psychological Science; Cortex; Conservation Biology*, *The Journal of Social Psychology* to name but a few. In institutions, proposed solutions include basing selection and promotion criteria on things other than publication metrics known to reward quantity over quality; creating departmental cultures of data sharing through training, example and policy, and being mindful of introducing new incentive structures that are current at odds with scientific integrity.

TOWARDS METHODOLOGY AND STATISTICS EDUCATION THAT IS INFORMED BY META-SCIENCE

As important as these structural solutions are, pedagogic solutions must run in parallel if science is to adequately meet reproducibility challenges. Firstly, because new content is required to prepare the next generation of scientists for a new set of expectations and requirements (e.g., training in data management, open science). Secondly, because the meta-science (meta-research) program that has grown up around the reproducibility crisis has identified a plethora of misconceptions and misunderstandings that require direct confrontation in the classroom.

Meta-science studies, amongst other things studies researchers' understanding of methodological and statistical processes; the completeness and transparency of their reporting practices; and the effectiveness of different structural and behaviour change interventions. It has

In M. A. Sorto, A. White, & L. Guyot (Eds.), *Looking back, looking forward. Proceedings of the Tenth International Conference on Teaching Statistics (ICOTS10, July, 2018), Kyoto, Japan.* Voorburg, The Netherlands: International Statistical Institute. iase-web.org

identified misconceptions and misunderstandings that would benefit from direct confrontation in the classroom. In this paper, we focus on two sets of meta-science findings. The first is the prevalence of Questionable Research Practices (QRPs) that have a direct impact on the reproducibility of published literature. The second relate to misconceptions about replication, including where information about replication comes from and the broader role replication plays in science. We discuss how teaching replication in a systematic way could help address both of these problems.

Questionable Research Practices

Questionable Research Practices (QRPs) include including cherry picking statistically significant results, *p* hacking, and hypothesizing after the results are known (HARKing). The widespread use of QRPs by researchers has been documented through self-report surveys (John et al 2012; Agnoli et al 2017; Fraser et al 2018). Rates across countries and disciplines are largely similar. For example:

- The proportion of researchers who had at least once withheld from publication variables that were not statistically significant (i.e., cherry picked results): 48% of US psychologists (John et al 2012), 63% of Italian psychologists (Agnoli et al 2017), 64% of ecology researchers and 64% of evolutionary biologists (Fraser et al 2018).
- The proportion of researchers who had at least once collected more data after first inspecting whether the results had reached statistical significance (i.e., a form of *p* hacking): 53% of US psychologists (John et al 2012), 56% of Italian psychologists (Agnoli et al 2017), 40% of ecology researchers and 51% of evolutionary biologists (Fraser et al, in prep).

The proportion of researchers who had at least once reported an unexpected result as though it had been predicted all along (i.e., HARKing): 37% of US psychologists (John et al 2012), 27% of Italian psychologists (Agnoli et al 2017), 49% of ecology researchers and 54% of evolutionary biologists (Fraser et al 2018).

Such practices have been directly implicated in the low rates of reproducible results uncovered by recent large scale replication studies in psychology and other disciplines. QRPs such as reporting only the subset of dependent/response/outcome variables or experimental conditions that reached statistical significance can inflate the false positive error rate of the research literature. Simons et al (2011) demonstrated this with simulated results, and warned of 'researcher degrees of freedom' in experimental reports, including failing to report the sampling stopping rule. This has been further demonstrated in an ecology and evolution context by Forstmeier et al (2017).

These and other similar demonstrations have a place in the statistics classroom. Many QRPs require direct confrontation. Students need to be not only warned of the potential damage they can cause to the scientific literature, but armed with scripts they can use to further confront colleagues and supervisors they may encounter in the workplace or in graduate studies who encourage them to engage in these practices.

Misconceptions about replication

Barnett et al (2018) found that whilst a majority of ecologists agreed that 'replication plays a crucial role in science and there should be more funding for it' (63%), they also grossly overestimated the amount of replication that currently takes place in the discipline. For example, the median estimate of the n=423 ecologists in Barnett et al's sample was that 15% of results in published literature are replicated, whereas literature surveys suggest it is likely to be far less. In open ended survey and interview responses, roughly a third of researchers of Barnett et al's sample suggested that if original experiments/studies expressed sufficient statistical information, the need for replication experiments/studies was diminished. This betrays a misunderstanding of the role replication plays in science and the scope of inferential statistics. A much smaller but still concerning number (~3%) explicitly expressed that if a single study had sufficient statistical power, replications of that study were completely redundant.

Lai et al (2010) found that researchers in psychology, medicine and statistics grossly overestimate the information *p* values provide about replication. Researchers were effectively asked: "Suppose you obtain $p = .02$ in an experiment, then replicate the experiment with new

samples. What *p* value might you obtain, and what interval has an 80% chance of including that replication *p*?" Under conservative assumptions the answer is a wide interval from .0003 to .30. Intervals returned by the n=360 researchers in Lai et al's sample corresponded to a 40% to 50% chance of including the replication *p* rather than the 80% stipulated in the survey question.

Further to the above, we know that many students and researchers have even more pronounced misconceptions about what *p* values tell us about the likelihood of replication. The 'replicability fallacy' is the false belief that a p value of .05 means that 95 times out of 100, the observed statistically significant difference will hold up in future investigations. In a now famous survey by Oakes (1986, p.79), 60% of researchers agreed with the following statement of the replicability fallacy: 'You have a reliable experimental finding in the sense that if, hypothetically, the experiment were repeated a great number of times, you would obtain a significant result on 99% of occasions.' In Haller and Krauss' (2002) replication of Oakes' study, 37% of methodology instructors, 49% of academic psychologists and 41% of psychology students agreed with the statement.

We suspect that very few (virtually no) methods and statistics classes explicitly or deliberately teach false interpretations of *p* values, or mislead students about the role of replication. Yet misconceptions and poor practices abound in the research community that is a result of this education. As we acknowledged in our introduction, explanations for QRPs are predominantly structural, but this should not imply that there isn't also need for an educative response. We have proposed that the QRPs that are directly implicated in irreproducibility, such as those discussed above, and misconceptions about replication and require direct confrontation in the classroom. There are many ways to do this. For example, Kalinowski et al (2008) outlined how teaching Bayes theorem can reduce misconceptions about *p* values, including the inverse probability fallacy. However, for the remainder of this paper we focus on a particularly promising method: the Collaborative Replications and Education Project (CREP).

TEACHING RELICATION: THE COLLABORATIVE REPLICATIONS AND EDUCATION PROJECT (CREP)

The Collaborative Replications and Education Project (CREP) is a crowd-sourced replication project explicitly designed to train undergraduate in open science practices (Grahe et al 2014; Grahe et al 2018). Instructors and students are invited to choose from a list of high impact psychology studies that are feasible for undergraduates to complete in a semester as a project for a research methods course. The CREP instructions and findings are all hosted on an Open Science Framework (OSF) project page (https://osf.io/wfc6u/wiki/home/) allowing for easy dissemination and tracking across wide geographic areas. Though it is currently only available in psychology, the procedures are easily transferable to other disciplines and with efforts underway to expand this to ecology and evolution (by the authors of this paper).

The procedure involves students contacting the CREP team and claiming a study to replicate. The students then "fork" the OSF page for the claimed study so that their work is permanently connected to the larger CREP project. They populate their own OSF page with a description of the research team and purpose of the project, the materials, a video of their procedure, and evidence of ethics approval. At this point the students submit their project for review by a CREP team including an Executive Reviewer, a student administrator, and up to more 2 faculty reviewers. The CREP team evaluates the project to make sure it matches the originally published work as closely as possible. The Executive Reviewer provides feedback to the students in a form similar to an editorial decision letter either approving the project for data collection or requesting revisions (ranging from minor to major). After the project is approved by the CREP team, students "register" their page (permanently freezing it in time) and begin data collection. Upon completion, the CREP team reviews the project again to evaluate the data and analyses components. Again there is a decision letter to help students present their research in the most transparent methods. Upon approval, students once again "register" their page, and are "certified" as a CREP study.

To date, over 350 students at 50 institutions started 114 projects. Though many never succeed in passing the second review, approved studies to date are present in one manuscript in

press (Leighton et al. 2018), and two more in preparation for submission (Ghelfi et al. 2018; Wagge et al. 2018), and there are too many student conference presentations to list here.

Contrasting this process to the typical student project highlights the pedagogical benefits that the CREP provides. Instead of students developing their own studies, students learn to conduct well-designed high impact studies. Instead of poorly developed hypotheses, students must understand theoretically meaningful hypotheses. Instead of creating unreliable measures without the time or material resources for pre-testing, they master and apply experimental materials and procedures that passed rigorous peer review standards of the high impact journals. Instead of completing studies that will be shared with the instructor, maybe the class, and sometimes at a conference; students work through the CREP review process that embodies professional experiences with studies intended to contribute to scientific literature.

Beyond the distinction between the type of studies and the intended outcomes, students learn Open Science principles by completing the process. They learn about the importance of replication and the risks of questionable research practices, and they also learn about solutions. With incremental steps, students learn how to use the OSF to present their ideas and research transparently. They learn how to organize pages to present their hypotheses and study plan (Preregistration), their materials and procedure (Open Materials), and their data and analyses (Open Data). They learn about the importance of replication and to temper their interpretations because we do not endorse conclusions about findings until there are sufficient replications

While all undergraduate research experiences are valuable learning experiences, and more are better (Taraban & Logue, 2012), the CREP employed replications for student projects is intended to augment the typical learning experience so that students can get even more from their experience as others suggested previously. In part, we intend to create greater motivation to do research; our common recruiting statement to students considering a CREP study is “Even though no one can guarantee results are publishable, at least three other PhDs will look at this research.” Further, we assert that learning to administer professionally developed research requires deeper learning than developing student level original projects.

The CREP currently exists in psychology with efforts to expand to ecology, but the model can be easily and cheaply extended into many other disciplines. A relatively small group of similarly motivated volunteers can devise a model for study selection, invite their own and others’ students to replicate those studies, and employ the same quality control checks to collect scientifically valid samples to be collated for dissemination. When the CREP began, it offered small monetary research awards for completed studies sponsored by Psi Chi and the Center for Open Science, but has continued successfully recruiting even after those funds expired. Otherwise, the CREP has operated completely on good will of volunteers, both faculty and students. It provides a lost cost example of high impact teaching practices.

CONCLUSION

Involving undergraduates in crowd-sourced replication studies has genuine pedagogic benefits (Frank & Saxe 2012; Grahe et al 2012) and provides an opportunity for directly confronting some particularly damaging misconceptions about the role of replication in science, and the role of inferential statistics. The CREP project is an established system for teaching research through conducting replications. It provides not only benefits for students and their instructors, but cumulative contributions can also help science as demonstrated by recent papers summarizing CREP findings (Leighton et al.; Ghee et al.; Wagge et al.). And most important, the CREP provides a model for other disciplines with plans underway to trial it in undergraduate ecology teaching and opportunity for motivated instructors in disciplines beyond

REFERENCES

Agnoli F, Wicherts JM, Veldkamp CLS, Albiero P, Cubelli R. Questionable research practices amongst italian research psychologists. (2017). *PLoS ONE, 12,* e0172792.

Fidler, F., Chee, Y.E., Burgman, M.A., McCarthy, M.A. & Gordon, A. (2017). Meta-research for evaluating reproducibility in ecology and evolution. *Bioscience, 67,* 282-289.

Forstmeier W, Wagenmakers EJ, Parker TH. (2017) Detecting and avoiding likely false-positive findings--a practical guide. *Biological Review, 92,* 1941–1968.

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

Freedman LP, Cockburn IM, Simcoe TE. (2015).The economics of reproducibility in preclinical research. *PLOS Biology, 13,* e1002165/

Ghelfi, E., Christopherson, C.D., Fischer, M.A., Legate, N., Lenne, R., Urry, H., Wagemans, F. M. A., Wiggins, B., Barrett, T., Glass, M., Guberman, J., Hunt, J., Issa, N., Paulk, A., Peck, T., Perkinson, J., Sheelar, K., Theado, R., Turpin, R. (2018). The influence of gustatory disgust on moral judgement: A pre-registered multi-lab replication. *Manuscript in preparation.*

Grahe, J. E., IJzerman, H., Brandt, M., Cohoon, J (March, 2014). Replication education. *APS Observer*. https://www.psychologicalscience.org/observer/replication-education

Grahe, J. E., Brandt, M. J., Wagge, J., Legate, N., Wiggins, B. J., Christopherson, C. D., … Baciu, C. (2018, February 22). Collaborative Replications and Education Project (CREP). http://doi.org/10.17605/OSF.IO/WFC6U

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

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

Haller, H., & Krauss, S. (2002). Misinterpretations of significance: A problem students share with their teachers? *Methods of Psychological Research, 7,* 1-20.

Kalinowski, P., Fidler, F., & Cumming, G. (2008). Overcoming the inverse probability fallacy: A comparison of two teaching interventions. *Methodology, 4, 152-158*

Leighton, D. C., Legate, N., LePine, S., Anderson, S. F., & Grahe, J. (in press). Self-esteem, self-disclosure, self-expression, and connection on Facebook: A collaborative replication meta-analysis. *Psi Chi Journal of Psychological Research.*

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*. 349(6251):aac4716. pmid:26315443.

Oakes, M.W. (1986). *Statistical inference: a commentary for the social and behavioural sciences.* Chichester, U.K: J. Wiley & Sons, Inc.

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

Simmons, J., 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.

Taraban, R., & Logue, E. (2012). Academic factors that affect undergraduate research experiences. *Journal of Educational Psychology, 104,* 499-514.

Wagge, J. Johnson, K., Meltzer, A., Baciu, C., Banas, K., Nadler, J. T., IJzerman, H., & Grahe, J. E. (2018). Elliott et al.'s (2011) "Red, Rank, and Romance" effect: A meta-analysis of CREP replications. *Manuscript in Preparation.*