Toward Greater Reproducibility of Undergraduate Behavioral Science Research

Bruce Evan Blaine, *St. John Fisher College*

Abstract

Reproducibility crises have arisen in psychology and other behavioral sciences, spurring efforts to ensure research findings are credible and replicable. Although reforms are occurring at professional levels in terms of new publication parameters and open science initiatives, the credibility and reproducibility of undergraduate research deserves attention. Undergraduate behavioral science research projects that rely on small convenience samples of participants, overuse hypothesis testing for drawing meaning from data, and engage in opaque statistical computing are vulnerable to producing nonreproducible findings. These vulnerabilities are reviewed, and practical recommendations for improving the credibility and reproducibility of undergraduate behavioral science research are offered.

Keywords: *data analysis, reproducibility, research methods, social sciences, statistics, undergraduate research*

doi: 10.18833/spur/3/1/2

Recent events in academic psychology, including the overturning of several "textbook" psychological effects, a widely reported replication project of prominent findings that returned disappointing results, and evidence of *p*-hacking and other questionable scientific practices, have created a credibility crisis in that field (Carter et al. 2015; John, Loewenstein, and Prelec 2012; Open Science Collaboration et al. 2015; Wagenmakers et al. 2016). Alarm bells around research credibility have been sounded in other behavioral science fields too, including economics (Ioannidis, Stanley, and Doucouliagos 2017; Necker 2014) and management (Bergh et al. 2017).

According to a National Science Foundation committee on replicability in science, reproducibility is a minimum necessary condition for a research finding to be credible and informative (Bollen et al. 2015). However, the concept of reproducibility can refer to any number of ideal goals, including transparency in research practices, replicating the results of past studies, systematic and detailed reporting of the details of research design, correcting a publication system that is biased toward novel and provocative research findings, and the proper use and interpretation of significance testing in data analysis. Goodman, Fanelli, and Ioannidis (2016) remind the researcher that the various meanings of reproducibility should not become an end in themselves, but rather move the researcher toward the ultimate goal—which is that the claims the researcher makes based on scientific research are, in fact, true.

In an influential paper, Ioannidis (2005) presented a framework for estimating the proportion of evidence-based claims that are actually true among published papers. He concluded that up to 50 percent of published claims are untrue, due to the presence of one or more of the following factors: low-power studies, *p*-hacking and other forms of bias driven by the goal of achieving statistically significant results, the number of other studies that exist on the same research question, and finally the base rate of true relationships to total relationships in a given field of research. Each of these factors affects a study's *positive predictive value* (PPV), which is the post-study probability that its claims, based on achieving statistical significance, are true. Although undergraduate research is seldom published, typical undergraduate research in the behavioral sciences has characteristics that predict low PPV, including small sample size, investigation of small-effect relationships, exploratory analyses that produce unpredicted research findings, and flexibility in design, measurement, and analytic methods. Consequently, it must be acknowledged that many undergraduate student research projects deliver findings that are probably not true.

Good science begins with reproducibility (Bollen et al. 2015), and helping undergraduate researchers pursue designs and analytic methods that increase the reproducibility of their research is an important element of the research experience. This can occur at the level of the student project or at the program level. At the project level, students who are mindful of reproducibility and include it in the goals of the research experience will do better research. As they understand that their project's conclusions are credible, and why they are, students may present and defend their research better. They may also read others' research claims more critically. At the program level, reproducibility guidelines or requirements for student research can have also desirable ripple effects. They encourage students to formulate better research problems and promote greater connection with the community of researchers on their particular question. Reproducibility requirements for student research also tie into curricular goals where research methods, data analysis, and statistical computing are concerned.

The Council on Undergraduate Research exists to "support and promote high-quality undergraduate student-faculty collaborative research" (CUR 2018). In that spirit, this paper identifies and explains four threats to the reproducibility of undergraduate behavioral science research and offers to faculty some practical recommendations and workarounds to improve overall reproducibility of student research. Faculty mentors of student research face a wide range of constraints and limitations, within which some of the following recommendations are not reasonable or attainable. The hope is that this article will, if nothing else, encourage small steps toward greater reproducibility. The analysis offered herein focuses primarily on independent (e.g., capstone, honors) research done under the supervision of a faculty mentor or advisor, although the recommendations have implications for course-related research. Finally, although many factors can affect reproducibility, including methodological and procedural aspects of research, this paper focuses on the quantitative and statistical elements of reproducibility.

Nonprobability Sampling

Most published research articles in psychology and related behavioral science fields use convenience samples of undergraduate students as data sources (Arnett 2008; Hanel and Vione 2016; Henrich, Heine and Norenzayan 2010; Peterson 2001). Convenience samples produce findings with low external validity, but the concern here is how nonprobability sampling undermines reproducibility. First, convenience samples represent unknown populations. Sampling from unknown populations makes parameter estimation less reliable, inasmuch as parameter estimates of different populations should not be expected to agree. For example, does a sample consisting of volunteers from a pool of students taking the introductory psychology course represent the psychology major population, the liberal arts student population, or some other population?

Second, convenience samples are often produced by an unknown sampling method. Convenience sampling (e.g., allowing research participants to sign up for a study) draws participants based on a mix of factors such as availability, interest in the research topic, coercion (e.g., course requirement), incentive (e.g., gift card lottery), and more. Independent of the population issue, sampling methods that are determined by a set of unknown and nonrandom factors cannot be expected to generate samples whose parameter estimates agree.

Third, nonrandom samples violate the i.i.d. assumption (wherein random variables are assumed to be independent and identically distributed) underlying most parametric statistical procedures. This assumption is crucial to the accuracy of normal-theory inference and when violated will bias estimates of standard errors of statistics used in data analysis. For researchers using parametric inferential procedures such as analysis of variance or least squares regression, the downstream effect of the biasing of standard errors is inaccurate p values and decisions based on p values. The combination of convenience sampling and normal theory statistical methods undermines the credibility of findings that are p-value based.

Peterson and Merunka (2014) investigated the ability of student convenience samples to produce reproducible results by having faculty at 49 business schools from across the United States administer a survey to a convenience sample of students at their school. The survey measured attitudes toward business ethics and capitalism; basic demographic variables were also measured. The results showed wide variability in the means and variances on each attitude measure across samples. To simulate the process of replicating a sample finding with an "equivalent" sample, the researchers compared each sample mean on the business ethics scale with every other sample mean, generating a set of 924 pairwise comparisons. Subjected to standard independent samples t tests, 31 percent of the comparisons achieved statistical significance ($p < .05$). Similar heterogeneity of findings was observed in the tests of group differences (using gender and religiosity as dichotomous grouping variables) in business ethics attitudes, including substantial variability in the direction of the group difference. Peterson and Merunka's study shows that even when convenience samples are drawn from an explicit population of interest (business school undergraduates) and use identical measures, research findings vary widely in both sign and magnitude.

Inasmuch as many student researchers and departments depend on participant pools and convenience samples for research participants, several methods can help students do more reproducible research without abandoning those key resources. First, students can and should do probability sampling for their studies, even if the population of interest (and available sampling frame) is narrow and restricts generalizability. For example, students can randomly sample from a sampling frame consisting of all students in all sections of a course, and then contact and schedule those students for the study. In this scenario the population of interest is much clearer (e.g., all introductory sociology students in public, four-year institutions), and the study will generate reasonably accurate parameter estimates of that population, provided that participant nonresponse does not bias the sample.

Second, student researchers should be encouraged to replicate their studies, if participants are available and if time permits. Recent large-scale efforts to replicate prominent psychology papers and findings have heightened awareness around the importance of replication to scientific credibility (Bohannon 2015; Klein et al. 2018; Open Science Collaboration et al. 2015). Similar efforts have emerged in economics (Duvendack, Palmer-Jones, and Reed 2015) and sociology (Freese and Peterson 2015). McShane and Böckenholt (2018) recommend single-paper meta-analysis, in which findings from the original study and its conceptual replication(s) are synthesized. Although this might be a lofty goal for typical undergraduate research studies, a meta-analysis of two small, low-powered studies is able to reveal effects that one or even both studies are unable to detect. Lastly, short of replication, students can be encouraged to do a pre-study power analysis, and design and conduct as powerful a study as resources and circumstances allow.

If student researchers replicating their own research is not feasible, there is value in students replicating published studies and in the process learning about open science and reproducibility. Numerous initiatives now support course-based and other student replication projects, such as the Collaborative Replications and Education Project (Open Science Framework 2013); Registered Replication Reports (Association for Psychological Science n.d.); and the Replication Network (n.d.). These initiatives provide online platforms for preregistration of studies, sharing protocols, data and findings, and collaboration with other researchers and labs. Furthermore, pedagogical models are emerging for instructors that confront the practical challenges of doing course-based or lab-based replication projects (Frank and Saxe 2012; Grahe, Guillaume, and Rudmann 2013; Hawkins et al. 2018; Janz 2016).

Third, students can consider secondary data sources for their research, such as in state or federal agency surveys or other secondary data sources (Sautter 2014). Most government surveys use sophisticated probability sampling methods and weighting schemes that cannot be achieved in undergraduate research. Other advantages of using state or federal survey data for research are the generally high quality of measurement, documentation, and relative ease of gaining permission and access to the data. The main disadvantage of using secondary data for research is that the study or survey may not contain measures of the variables of interest. If proxy variables can be identified, however, the benefits of doing research using secondary data are compelling.

Low Positive Predictive Value

Cohen (1994) observed that most studies in psychology do not have enough statistical power to detect the effects they are trying to detect. Statistical power is related to a study's positive predictive value (PPV), which as mentioned earlier is the post-study probability that a study's claims, based on achieving statistical significance, are true. Even when optimistic values for Ioannidis's PPV equation parameters R and u (see Ioannidis 2005) are substituted, low-power exploratory studies have PPVs in the 5 to 10 percent range. This means that a claim based on formal significance testing in a low-power exploratory study is very unlikely to be true. Keep in mind that a study is still considered exploratory if it tests a priori hypotheses but reports other unpredicted significant findings discovered in the data, a common practice in behavioral science research.

The best evidence for making credible truth claims are large-scale controlled experiments (random controlled trials, or RCTs) or meta-analyses of RCTs. Obviously, a large-scale controlled experiment is not a practical undergraduate research option. Meta-analysis is a more reasonable option for undergraduate research, and recommends itself for several reasons (Chan and Arvey 2012). First, the vast primary research literatures that have accumulated across the behavioral sciences mean that most research questions posed by undergraduate students have been examined by a great many published studies. Moreover, these literatures are searchable and retrievable through library tools available at most institutions. Second, meta-analytic research engages student researchers in the primary research around their question and invites them to distinguish better from lesser quality research, examine evidence in the form of treatment or relationship effect sizes, and synthesize the research evidence across a set of conceptual replication studies. Third, a competent meta-analysis can be done with the statistical background of an undergraduate statistics course. If students understand analysis of variance and least squares regression, they can master the basic statistical procedures for meta-analysis. Resources abound for making meta-analysis a more accessible research option for students (APA Science Student Council 2008; Field and Gillett 2010). Fourth, a meta-analysis of low-power studies has a far higher PPV, and thus more credible truth claims, than could be achieved by a single study.

Recommendations for improving the PPV of undergraduate student research projects merely reiterate Cumming's (2012) recommendations for psychological science in his "new statistics" framework: pursue more meta-analytic thinking and effect size estimation, avoid questionable data analytic and data reporting practices, move away from null hypothesis significance testing (see next section), and encourage replication. Admittedly, encouraging meta-analysis and mentoring students in meta-analytic research imposes some demands on faculty that may not be realistic. However, short of that, several small steps can improve the credibility (through higher PPV) of undergraduate research studies.

First, discourage the reporting of unpredicted significant findings from research data, unless those findings are then replicated in a follow-up study. Discovered significant findings capitalize on chance, and thus such findings, if reported, should be presented with an appropriate level of untrustworthiness. Second, given that most undergraduate projects are single-study primary research, encourage students to think about the size of the effect or relationship the study is trying to detect, get estimates of the effect from the literature, and then design a study with enough power to detect that effect. Third, encourage students to prepare, follow, and make public detailed methodological and data analytic plans for their study. This need not mean formal preregistration of a student's project. There are many more modest and achievable ways to allow students to make their data and data analytic work open to a broader community, such as a department or school page for posting research protocols or through open portfolios of student research. Research plans reduce bias—defined here as all the subtle deviations, accommodations, and changes to a study's method and data analysis that result in better findings. Public research plans and protocols encourage replication and credibility.

Overemphasized Null Hypothesis Testing

The limitations of null hypothesis significance testing (NHST) are well established and supported by a critical literature going back 50 years. Over that period there have been persistent calls in the behavioral sciences to reform the conventions around NHST, ranging from supplementing NHST with other inferential methods to limiting its use to abandoning it altogether (see Cohen 1994; Kline 2004; Krantz 1999; Wilkinson 1999). Recently, Cumming (2014) has put the NHST issue front and center in his new statistics framework, arguing compellingly that NHST has no place in research that strives to be credible and reproducible.

One of the problems in helping students become less dependent on NHST for doing statistical inference is that the method is still dominant in undergraduate statistics textbooks written for the behavioral science audience. A recent survey of undergraduate sociology programs found that most programs (67 percent) required one statistics course, but a large minority of programs (27 percent) did not require a statistics course (Delia Deckard 2017). It did not matter much whether the undergraduate degree was obtained from a liberal arts college, regional university, or large research university. These findings suggest that students may not have opportunities to learn alternative procedures to NHST, and this presents program-level challenges for addressing this part of the research reproducibility problem.

Nevertheless, for students and faculty mentors who want to move beyond the limitations of NHST and thereby improve the reproducibility of the research, two general goals should be pursued: to emphasize parameter estimation and effect size statistics over hypothesis testing, and to use more accurate significance testing procedures. Here are some specific, achievable recommendations to move toward each goal. First, significance tests are an obstacle to cumulative knowledge inasmuch as they seem to provide a precise and authoritative "answer" to a question. Confidence intervals of relationships or treatment effects, by contrast, display the probable range of "answers" that a study could easily have generated, helping students to see both the estimates (reflecting the size of the effect) and the imprecision in their study's ability to answer a question (Cumming and Fidler 2009; Smithson 2011).

Confidence intervals also allow the researcher to conduct strong (in Cohen's [1994] terms, non-nil null) hypothesis tests, such as whether a correlation is reliably larger than some crud factor in a particular field (Meehl 1997). Finally, even within single studies, confidence intervals encourage the student researcher to accumulate evidence for or against a hypothesis (e.g., estimates of a treatment effect on multiple outcomes, or in different subgroups of participants)—which is the very essence of the meta-analytic thinking of Cumming (2012, 2014). By quantifying the amount of an observed relationship or effect, effect size statistics convey clinical or practical significance, which is far more important than statistical significance. To that end, unstandardized effect size statistics (e.g., unstandardized regression coefficient) are highly recommended for reporting because they are based on meaningful measurement metrics, they are simpler to explain, and they are more transparent to practical significance (Pek and Flora 2018).

Second, with regard to the goal of doing better significance testing, it is worth noting that the problems associated with NHST are not in the method itself, but in its misapplication and misinterpretation of *p* values. The null hypothesis test does one thing very well—it helps the researcher discredit the sampling error hypothesis—and should not be discarded because people misuse it. With that said, parametric statistics are commonly used in behavioral research for hypothesis testing. The problem for research credibility and reproducibility is that parametric procedures only work well when their assumptions are met in the data. The widespread reliance on nonprobability samples, addressed earlier, violates one parametric assumption. But considerable evidence shows that nonnormality and heterogeneity of variance are more the rule than exceptions to the rule in behavioral research and that researchers rarely test for violations to parametric assumptions (Grissom 2000; Keselman et al. 1998; Ruscio and Roche 2012). When their assumptions about the data are violated, parametric inferential tests generate inaccurate *p* values and Type I error rates, contributing to less credible statistical decisions. Fortunately, for every parametric test, there is an equivalent nonparametric or resampled test procedure. These procedures make far fewer assumptions about the data, perform better with small samples, and thus help the researcher deliver more accurate and reproducible statistical inferences. In addition, most nonparametric and resampling-based inferential procedures can be done in SPSS and other statistical packages used in the behavioral sciences.

Opaque Statistical Computing

Single-sample primary studies are often not reproducible (Bohannon 2015; Epstein 1980; Open Science Collaboration et al. 2015). One reason for this is opaque statistical computing: data analytic procedures and actions being unknown to anyone but the researcher. And, with menu-driven statistical software, even researchers can lose track of what they did to or with the data. Accordingly, a reproducible study requires that two things be made available to other researchers: the data from the original study and the documented data analytic operations used to analyze the data (Peng 2015). In addition to ensuring a study's replicability and thereby the credibility of its claims, these two provisions offer other benefits to the research community in any particular discipline. First, research articles omit many data analytic details (often necessarily, due to page limits) that are important for the credibility of their claims. Sharing data analytic decisions and protocol helps the audience better understand the findings and how the researcher arrived at them. Second, shared data analytic methods allow others to apply them to their data. As a consequence, innovations and new applications spread through the research community in ways that cannot be achieved through the vehicle of published papers (Open Science Framework n.d.).

As reviewed at the beginning of this article, current data analytic practices in the behavioral sciences are still largely inherited from an old irreproducibility model. Under that model, research data would be entered into a file, cleaned and prepared for analysis, and then analyzed using analytic procedures chosen from pull-down menus, from which some pieces of the output would find their way into the report. In this scenario, there is no way to know how the raw data were changed in the cleaning process or why. There is also no way to know how any of the analyses were done, what analyses were not done, or what analyses were done and not reported. To the extent these workflow details are not reported in the particular work product, research conducted under this model is not replicable, even if it is otherwise excellent in design, and thus its findings are not credible.

Happily, open science initiatives are pushing reform and change in many disciplines (Open Science Collaboration 2015). Two points must be made about the implications of open data and statistical analysis for both programs and mentoring student researchers. One, open science encourages the use of software tools that are built to accommodate sharing and collaboration; some of the most widely used include R, R Markdown, and Github, but there are many other tools, including the replication and open science digital communities discussed earlier (e.g., Open Science Framework n.d., Collaborative Project; Association for Psychological Science, n.d.; Replication Network n.d.). Faculty that are comfortable with SPSS or another statistics package need not worry; research credibility does not depend on learning R.

Two, research sharing and collaboration or replication projects may require working in a common computing environment and language (e.g., R), so undergraduate students participating in those projects will need to be introduced to those computing tools. Participating in open science initiatives, therefore, may impose some curricular and resource demands on a faculty mentor or a program that may not be realistic. Nevertheless, Bray and his colleagues' thoughts are worth considering: they argue that an open science framework for undergraduate research is much better for the students (Bray, Çetinkaya-Rundel, and Stangl 2014). They make the following points. First, when students analyze data using a code-based (compared to menu-driven) environment, they are much more "in conversation" with the data, learning from it and letting the data inform the analysis. This workflow mimics the draft-revision-final draft process in writing, and contributes to better reporting of the project. Second, data cleaning (i.e., missing data procedures, recodes, variable creation) involves many decisions and steps that are both crucial to the analysis and underreported in articles. Analysis transparency preserves all of those steps, in sequence, for others to see. Third, statistical programming files allow students to work collaboratively and, using a sharing tool like those mentioned earlier, work on the same file using the run-revise-rerun workflow mentioned above. Fourth, statistical programming files allow the analyst's thoughts and rationale to be inserted between code chunks in a way that conveys a vivid "story" of the analysis. Finally, this reproducible framework for student research project work enables faculty to evaluate students' work more precisely, because they can see not only what students did, but what their thoughts were, at each step of the analysis.

Conclusion

Open science initiatives have begun in psychology and other behavioral sciences (Höffler 2017; Novotney 2014), reforming professional practices to increase research integrity, credibility, and reproducibility. Predictably, graduate training programs are responding in kind, preparing researchers that practice more reproducible science. This article has offered some practical thoughts on quantitative and statistical practices that contribute to more credible and reproducible research, with the undergraduate behavioral science student (and department) in mind. Faculty advisers can help students do research that produces more credible claims by paying attention to issues like probability sampling, statistical power, PPV, alternatives to null hypothesis testing, and more transparent statistical computing. Acknowledging the constraints faced by faculty and programs, any small steps to increase student research credibility matter, because they promote research and statistical principles that are worth pursuing and that contribute to better science.

References

American Psychological Association (APA) Science Student Council. 2008. Accessed September 5, 2019. http://www.apa.org/science/about/psa/2008/04/ssc.aspx

Arnett, Jeffrey. 2008. "The Neglected 95%: Why American Psychology Needs to Become Less American." *American Psychologist* 63: 602–614. doi: 10.1037/0003-066x.63.7.602

Association for Psychological Science. n.d. Registered Replication Reports. Accessed September 5, 2019. https://www.psychologicalscience.org/publications/replication

Bergh, Donald D., Barton M. Sharp, Herman Aguinis, and Li Ming. 2017. "Is There a Credibility Crisis in Strategic Management Research? Evidence on the Reproducibility of Study Findings." *Strategic Organization* 15: 423–436. doi: 10.1177/1476127017701076

Bohannon, John. 2015. "Many Psychology Papers Fail Replication Test." *Science* 349(6251): 910–911. doi: 10.1126/science.349.6251.910

Bollen, Kenneth, John T. Cacioppo, Robert M. Kaplan, Jon A. Krosnick, and James L. Olds. 2015. *Social, Behavioral, and Economic Sciences Perspectives on Robust and Reliable Science: Report of the Subcommittee on Replicability in Science Advisory Committee to the National Science Foundation Directorate for Social, Behavioral, and Economic Sciences*. National Science Foundation, Arlington, VA. Accessed September 5, 2019. https://www.nsf.gov/sbe/AC_Materials/SBE_Robust_and_Reliable_Research_Report.pdf

Bray, Andrew, Mine Çetinkaya-Rundel, and Dalene Stangl. 2014. "Taking a Chance in the Classroom: Five Concrete Reasons Your Students Should Be Learning to Analyze Data in the Reproducible Paradigm." *Chance* 27(3): 53–56. doi: 10.1080/09332480.2014.965635

Carter, Evan C., Lilly M. Kofler, Daniel E. Forster, and Michael E. McCullough. 2015. "A Series of Meta-Analytic Tests of the Depletion Effect: Self-Control Does Not Seem to Rely on a Limited Resource." *Journal of Experimental Psychology, General* 144: 796–815. doi: 10.1037/xge0000083

Chan, MeowLan Evelyn, and Richard D. Arvey. 2012. "Meta-Analysis and the Development of Knowledge." *Perspectives on Psychological Science* 7: 79–92. doi: 10.1177/1745691611429355

Cohen, Jacob. 1994. "The Earth Is Round (p < .05)." *American Psychologist* 49: 997–1003. doi: 10.1037//0003-066x.49.12.997

Council on Undergraduate Research (CUR). n.d. *Constitution and Bylaws of the Council on Undergraduate Research, Article I, Section 2*. Accessed September 5, 2019. https://www.cur.org/assets/1/7/Constitution_and_Bylaws.pdf

Cumming, Geoff. 2012. *Understanding the New Statistics: Effect Sizes, Confidence Intervals, and Meta-Analysis*. New York: Routledge.

Cumming, Geoff. 2014. "The New Statistics: Why and How?" *Psychological Science* 25: 7–29. doi: 10.1177/0956797613504966

Cumming, Geoff, and Fiona Fidler. 2009. "Confidence Intervals: Better Answers to Better Questions." *Zeitschrift Für Psychologie/Journal of Psychology* 217: 15–26. doi: 10.1027/0044-3409.217.1.15

Delia Deckard, Natalie. 2017. "Statistics Education for Undergraduate Sociology Majors: Survey Findings across Institutions." *Numeracy* 10(2): Article 8. doi: 10.5038/1936-4660.10.2.8

Duvendack, Maren, Richard W. Palmer-Jones, and W. Robert Reed. 2015. "Replications in Economics: A Progress Report." *Economics in Practice* 12: 164–191.

Epstein, Seymour. 1980. "The Stability of Behavior: II. Implications for Psychological Research." *American Psychologist* 35: 790–806. doi: 10.1037//0003-066x.35.9.790

Field, Andy P., and Raphael Gillett. 2010. "How to Do a Meta-Analysis." *British Journal of Mathematical and Statistical Psychology* 63: 665–694. doi: 10.1348/000711010x502733

Frank, Michael C., and Rebecca Saxe. 2012. "Teaching Replication." *Perspectives on Psychological Science* 7: 600–604. doi: 10.1177/1745691612460686

Freese, Jeremy, and David Peterson. 2015. "Replication in Social Science." *Annual Review of Sociology* 43: 147–165. doi: 10.1146/annurev-soc-060116-053450

Goodman, Steven N., Danielle Fanelli, and John P. A. Ioannidis. 2016. "What Does Research Reproducibility Mean?" *Science Translational Medicine* 8(341): 341ps12. doi: 10.1126/scitranslmed.aaf5027

Grahe, Jon, Esther Guillaume, and Jerry Rudmann. 2013. "Students Collaborate to Advance Science: The International

Situations Project." *CURQ on the Web* 34(2): 4–9. Accessed September 5, 2019. https://www.cur.org/assets/1/23/Winter2013_v34.2_Grahe.Guilaume.Rudmann.pdf

Grissom, Robert J. 2000. "Heterogeneity of Variance in Clinical Data." *Journal of Consulting and Clinical Psychology* 68: 155–165. doi: 10.1037//0022-006x.68.1.155

Hanel, Paul H. P., and Katia C. Vione. 2016. "Do Student Samples Provide an Accurate Estimate of the General Public?" *PLOS ONE* 11(12): 1–10. doi: 10.1371/journal.pone.0168354

Hawkins, Robert X. D., Eric N. Smith, Carolyn Au, Juan Miguel Arias, Rhia Catapano, Eric Hermann, Martin Keil, et al. 2018. "Improving the Replicability of Psychological Science through Pedagogy." *Advances in Methods and Practices in Psychological Science* 1: 7–18. doi: 10.1177/2515245917740427

Henrich, Joseph, Steven J. Heine, and Ara Norenzayan. 2010. "The Weirdest People in the World?" *Behavioral and Brain Sciences* 33: 61–135. doi: 10.1017/S0140525X0999152X

Höffler, Jan H. 2017. "Replication and Economics Journal Policies." *American Economic Review* 107(5): 52–55. doi: 10.1257/aer.p20171032

Ioannidis, John P. A. 2005. "Why Most Published Research Findings Are False." *PLOS Medicine* 2(8): e28. doi: 10.1371/journal.pmed.0020124

Ioannidis, John P. A., T. D. Stanley, and Hristos Doucouliagos. 2017. "The Power of Bias in Economics Research." *Economic Journal* 127: F236–F265. doi: 10.1111/ecoj.12461

Janz, Nicole. 2016. "Bringing the Gold Standard into the Classroom: Replication in University Teaching." *International Studies Perspectives* 17: 392–407. doi: 10.1111/insp.12104

John, Leslie K., George Loewenstein, and Drazen Prelec. 2012. "Measuring the Prevalence of Questionable Research Practices with Incentives for Truth Telling." *Psychological Science* 23: 524–532. doi: 10.1177/0956797611430953

Keselman, H. J., Carl J. Huberty, Lisa M. Lix, Stephen Olejnik, Robert A. Cribbie, Barbara Donahue, Rhonda K. Kowalchuk, et al. 1998. "Statistical Practices of Educational Researchers: An Analysis of Their ANOVA, MANOVA, and ANCOVA Analyses." *Review of Educational Research* 68: 350–386. doi: 10.3102/00346543068003350

Klein, Richard, Michelangelo Vianello, Fred Hasselman, Byron G. Adams, Reginald B. Adams Jr., Sinan Alper, Mark Aveyard, et al. 2018. "Many Labs 2: Investigating Variation in Replicability across Sample and Setting." *PsyArXiv Preprints*. doi: 10.31234/osf.io/9654g

Kline, Rex B. 2004. *Beyond Significance Testing: Reforming Data Analysis Methods in Behavioral Research*. Washington, DC: American Psychological Association. doi: 10.1037/10693-000

Krantz, David H. 1999. "The Null Hypothesis Testing Controversy in Psychology." *Journal of the American Statistical Association* 44: 1372–1381. doi: 10.1080/01621459.1999.10473888

McShane, Blakeley B., and Ulf Böckenholt. 2018. "Want to Make Behavioural Research More Replicable? Promote Single-Paper Meta-Analysis." *Significance* 15(6): 38–40. doi: 10.1111/j.1740-9713.2018.01214.x

Meehl, Paul. 1997. "The Problem Is Epistemology, Not Statistics: Replace Significance Tests by Confidence Intervals and Quantify Accuracy of Risky Numerical Predictions." In *What If There Were No Significance Tests?*, ed. Lisa Harlow, Stanley Mulaik, and James Steiger, 393–425. Mahwah, NJ: Erlbaum.

Necker, Sarah. 2014. "Scientific Misbehavior in Economics." *Research Policy* 43: 1747–1759. Accessed March 1, 2018. doi: 10.1016/j.respol.2014.05.002

Novotney, Amy. 2014. "Reproducing Results." *Monitor on Psychology* 45(8): 32.

Open Science Collaboration. 2015. "Estimating the Reproducibility of Psychological Science." *Science* 349(6251): aac4716. doi: 10.1126/science.aac4716

Open Science Framework . n.d. Accessed January 31September 5, 2019. https://osf.io/

Open Science Framework. 2013. "Collaborative Replications and Education Project (CREP)." Accessed September 5, 2019. doi: 10.17605/OSF.IO/WFC6U

Pek, Jolynn, and David Flora. 2018. "Reporting Effect Sizes in Original Psychological Research: A Discussion and Tutorial." *Psychological Methods* 23: 208–225. Accessed March 1, 2018. doi: 10.1037/met0000126

Peng, Roger. 2015. "The Reproducibility Crisis in Science: A Statistical Counterattack." *Significance* 12(3): 30–32. doi: 10.1111/j.1740-9713.2015.00827.x

Peterson, Robert A. 2001. "On the Use of College Students in Social Science Research: Insights from a Second-Order Meta-Analysis." *Journal of Consumer Research* 28: 450–461. doi: 10.1086/323732

Peterson, Robert A., and Dwight R. Merunka. 2014. "Convenience Samples of College Students and Research Reproducibility." *Journal of Business Research* 67: 1035–1041. doi: 10.1016/j.jbusres.2013.08.010

The Replication Network: Furthering the Practice of Replication in Economics. n.d. Website. Accessed September 5, 2019. https://replicationnetwork.com

Ruscio, John, and Brendan Roche. 2012. "Variance Heterogeneity in Published Psychological Research: A Review and a New Index." *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences* 8: 1–11. Accessed March 1, 2018. doi: 10.1027/1614-2241/a000034

Sautter, Jessica M. 2014. "Secondary Analysis of Existing Data in Social Science Capstone Research." *CURQ on the Web* 34(4): 24–30.

Smithson, Michael. 2011. *Confidence Intervals*. Quantitative Applications in the Social Sciences, no. 07-140. Thousand Oaks, CA: Sage Publications.

Wagenmakers, E-J, Titia Beek, Laura Dijkhoff, and Quentin Gronau. 2016. "2016 Registered Replication Report: Strack, Martin,

& Stepper (1988).” *Perspectives on Psychological Science* 11: 917–928. doi: 10.1177/1745691616674458

Wilkinson, Leland. 1999. “Statistical Methods in Psychology Journals Guidelines and Explanations.” *American Psychologist* 54: 594–604. doi: 10.1037/0003-066x.54.8.594

Bruce Evan Blaine
St. John Fisher College, bblaine@sjfc.edu

Bruce Evan Blaine is an applied statistician and professor of statistics and data sciences in the Department of Mathematical and Computing Sciences at St. John Fisher College in Rochester, NY. His interests include meta-analysis, nonparametric statistics, and statistical methods used in the behavioral sciences. He teaches a range of courses, including Introduction to Data Science, Nonparametric Statistics, Predictive Analytics, Meta-Analysis, and Quantitative Research Methods. He is also an American Statistical Association–accredited Professional Statistician.® Parallel to his academic responsibilities, he consults with students and community clients on research design and statistical issues.