Research and Creative Activities Reproducibility Subcommittee

Report to OKED

April, 2016

Summary

There has been a growing interest over the last couple of years in the issue of reproducibility of scientific research, or more precisely the lack thereof. This interest was accelerated to a large extent by publication of the Reproducibility Project: Psychology in August, 2015. The current report, prepared by the Reproducibility Subcommittee of the Research and Creative Activities Senate committee, begins with a brief summary of that Project's findings, with an emphasis on using it to clarify the nature of the reproducibility crisis. In discussing some of the reasons for poor rates of reproducibility, we describe how problems with reproducibility are only the most visible symptom of systemic problems with many current scientific practices.

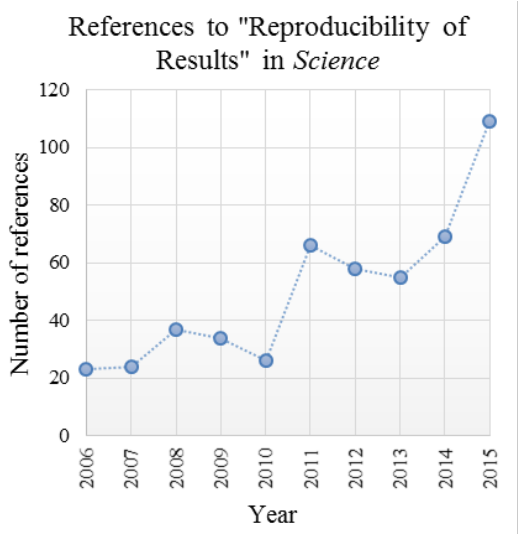
The report then turns to current efforts to address these problems. In fact, extensive resources are already available to facilitate the transition to a more transparent and open practice of science. We provide a brief introduction to some of these resources.

The report ends by describing the role that various entities within the scientific community might play in the transition toward best scientific practices. Specifically, we discuss the role of journals, professional organizations, funding sources, and finally, universities.

Contents

I) Introduction to the issue ↑

Over the past year or so, the scientific community's interest in reproducibility, or more accurately the lack thereof, has gone viral. For example, references to reproducibility in *Science* have markedly increased over the last 10 years (see figure)[1]. Likewise, PLOS published a collection of articles from their journals on the topic of meta-research, inspired by the reproducibility crisis.



In this report, reproducibility will refer to the ability to repeat an experiment under conditions as close to the original as possible and obtain the same qualitative pattern of results to an extent that cannot be attributed to chance. Many other terms, such as repeatability, reliability, and replication, are used to refer to similar concepts, although there may be subtle differences in meaning depending on discipline or context. For example, replication may refer to cases where the same qualitative pattern of results occurs when conditions are intentionally changed in order to broaden the scope of the findings. By these definitions, failures to replicate are often informative and advance understanding. Failures to reproduce results are always cause for concern. In the current report, we use these terms interchangeably (where appropriate, using the same term as the work being discussed).

Early meta-studies investigating reproducibility produced troubling results. In 2011, independent studies at Bayer (Begley & Ellis, 2012) and Amgen (Prinz, Schlange, & Asadullah, 2011) attempted to reproduce the results of many groundbreaking basic science studies. They achieved success rates of only 11% and 25% respectively. The ManyLabs project conducted a similar investigation, in which they attempted to reproduce 13 findings in Cognitive and Social Psychology in 36 independent labs around the world. Although they achieved a higher success rate, the results showed that results were not as reproducible as they should be.

It is often said that science is self-correcting. True to this claim, findings like these inspired action. Reproducibility is only the most visible (and arguably most alarming) symptom of problems with the current practice of science. The deeper problem is a serious lack of transparency and openness in science as it is currently practiced. The problems in biomedical research led the founding of the Global Biological Standards Institute in 2012, which "is dedicated to enhancing the quality of biomedical research by advocating best practices and standards to accelerate the translation of research breakthroughs into life-saving therapies." They developed the action plan "Reproducibility 2020," which aims to greatly improve the quality of preclinical biomedical research. In 2013, the Center for Open Science was founded to promote

[1] Number of references found using *Science* website search engine, search term "reproducibility of results," search time frame 1/1/2006 to 1/1/2016.

greater openness and transparency across all fields of science.

The good news is that the move toward open science should produce benefits much greater than simply increasing the rate of reproducibility. In addition to improving the quality of the scientific corpus, there are financial benefits as well. For example, it is estimated that the sharing of fMRI data through the OpenfMRI project, which currently contains only 42 data sets, has already produced a savings of $878,400 over five years (Gorgolewski et al., 2015).

Because most of the early work promoting open science practices has centered around the biomedical field and cognitive and social psychology, we will use examples from these domains; however, there is activity across a range of fields. Earth and Planetary Sciences seem particularly involved. In theoretical Physics, which is one of the most advanced fields in terms of open and transparent practices, the Laser Interferometer Gravitational-wave Observatory (LIGO), which recently published the highly publicized discovery of gravitational waves, operates an open science center.

The first major project undertaken by the Center for Open Science was the Reproducibility Project: Psychology (RP:P). It published its findings in Science in August 2015. A rough characterization of the findings is that only about 30% of the 100 studies that were repeated were reproduced (a more careful characterization is offered below).

Not surprisingly, the Reproducibility Project: Psychology has received significant media attention. Stories have appeared in a wide range of publications, from the New York Times to Buzzfeed. They have also ranged widely in perspective from balanced to alarmist. The Center for Open Science provides a more extensive collection of press clippings.

The second reproducibility project is the Reproducibility Project: Cancer Biology (RP:CB). If it fails to reproduce cancer research at any noteworthy level, which it almost certainly will, we can expect *much* more visible coverage in the popular press.

Because the Reproducibility Project: Psychology seems to have accelerated interest in the topic of reproducibility, it is a good place to start.

II) Reproducibility Project: Psychology ↥

Previous reports of poor replicability in medicine (Begley & Ellis, 2012; Prinz et al, 2011), along with descriptions of problematic practices and incentives in data analysis and publication within the field of psychology (for example, Ioannidis, 2005), inspired a group of psychologists (Open Science Collaboration, OSC) to conduct a large-scale project (the Reproducibility Project: Psychology; RP:P) aimed at deriving an initial estimate of reproducibility in psychological science. Three states of reality can produce failures to replicate: Differences between the original study and the attempted replication, a false positive in the original study, and/or a false negative in the replication study. Planners of the RP:P did their best to eliminate the first of these.

A) Method ↥

The original OSC collaborative team developed a standardized replication protocol that carefully specified processes for 1) choosing studies and key effects to be replicated; 2) contacting original authors to obtain study materials; 3) preparing a protocol and analysis plan

for the replication; 4) soliciting review of the replication plan from the original authors and other members of OSC and integrating their suggestions; 5) publicly registering the protocol; 6) carrying out the replication project; 7) writing the report; and 8) quality control for all aspects of the project. All replication materials, data, and reports, along with R code for replication analyses, is archived publicly.

Articles were selected those published in 2008 in three top-tier psychology journals: Psychological Science (publishes articles in all psychology subdisciplines), Journal of Personality and Social Psychology, and Journal of Experimental Psychology: Learning, Memory, and Cognition. This sampling frame was chosen to 1) enable comparisons across social and cognitive subdisciplines; 2) fit expertise of original collaborators; 3) be recent enough to obtain original study materials; 4) be old enough to have meaningful impact indicators; and 5) represent subdisciplines in which studies are relatively low cost (to increase feasibility of replication).

In all, 270 researchers contributed to the project. Study selection began using the first 20 articles from 2008 in each journal. Project coordinators helped replication teams match interests and expertise with appropriate studies from this pool until matching became too difficult, at which point another 10 articles were made available from one or more of the journals; this process was iterated. Project coordinators also recruited research teams to match specific articles. Of those articles tallied in the major replication report, 43 were cognitive and 57 were social-personality. Although many articles presented multiple studies and results, standardized criteria were used to choose for replication a single key result from a single experiment in each article. Original and replication results were converted to correlation coefficients (*r*) with confidence intervals (CI), which served as a common metric of effect sizes.

B) Results ↥

Reproducibility was evaluated by comparing the following indicators between the original and the replicated studies: statistical significance, p values, effect sizes, and subjective assessments of replication success by the research teams. Although there were some differences, all of these comparisons converged to suggest that reproducibility was low.

Statistical significance vs. non-significance. Ninety-seven of the original 100 effects were statistically significant; if all of these effects were true, 89 of the replications should have been significant, but only 35 were significant.

P-values. The mean *p*-value of the original studies was .028, whereas the mean *p*-value of the replicated studies was .302. The distribution of replication *p*-values argues against the explanation that these failures to replicate are explained by low statistical power.

Effect sizes. The original effect sizes (mean = .403, SD = .188) were larger than the replication effect sizes (mean = .197, SD = .257); 82.8% showed a stronger effect size in the original study. Original and replication effect sizes were positively correlated.

Subjective assessments of success. Of the 100 effects reported, research teams rated 39 of them as having successfully replicated the original effects.

Specific characteristics of the original studies and replications were assessed and tested as

potential influences on reproducibility. These included 1) original publishing journal; 2) importance of effect (as indicated by citation indices); 3) rated "surprisingness" of effect; 4) sample size, p value, effect size, and statistical power for both original and replicated effect; 5) original and replication teams' expertise and experience; 6) independently rated difficulty of conducting effective replication, and 7) self-assessed quality of replication effort.

Replication success was best predicted by the strength of the original findings, as indicated by statistical significance, low *p*-value, and large effect size. For example, 41% of original studies with $p < .02$ achieved $p < .05$ in the replication, whereas only 18% of those with $p > .04$ did so. Using statistical significance as the indicator of success, cognitive psychology effects were more reproducible than social psychology effects (50% vs. 25%), but this difference did not hold in the comparison of effect sizes. These findings may be explained by the relatively greater frequency of high-power within-subject designs in cognitive psychology. Characteristics of the original and replication research teams were not predictive of replication.

C) Comments on results from RP:P ↥

Despite using materials provided by the original authors, careful review in advance to ensure methodological similarity to the original study, and relatively high statistical power, most of the replication studies produced significantly weaker evidence than did the original studies. It is important to note, however, that failure to replicate does not mean that the original finding was a false positive; unknown factors may have influenced the results of the replication study, or it may have produced a false negative result. It is also important to note that successful replication speaks to the reliability of a finding, but not to its theoretical explanation. Alternative explanations may apply to both studies.

An independent re-analysis of 72 of the replication efforts using Bayesian statistics provided some interesting insights into why so many studies investigated in the RP:P failed to replicate (Etz & Vandekerckhove, 2016). Bayesian analysis allows one to evaluate the strength of evidence supporting the null hypothesis as well as the strength of evidence supporting the alternative hypothesis. None of the studies analyzed provided strong evidence (Bayes factor < 10) supporting the null hypothesis. However, 64% of the studies did not provide strong evidence for either the null hypothesis or the alternative hypothesis in either the original study or in the replication study. Based on this analysis, a very plausible conclusion one might draw is that sample sizes are too small in much of psychological research, which reduces the likelihood that significant effects are true effects. Naturally, this lowers the likelihood of reproducing such effects.

Furthermore, scientific communication in many fields, including psychology, is strongly biased toward the publication of 1) novel findings and 2) statistically significant results. Together, these two practices almost guarantee that replication studies will be few, and that those undertaken will show weaker results than the original, published studies. Therefore, it is likely that the high rate of failure to replicate in the RP:P reflects the fact that the published literature is the result of many "rolls of the dice" due to publication bias, while each finding from the RP:P is the result of a single "roll of the dice."

Pre-registered studies with high-quality methods and certainty of publication will reduce editorial bias that leads to the publication of primarily statistically significant results. Finding a balance between innovation and verification will be similarly important.

III) Why these problems with reproducibility? ↥

Several interrelated factors contribute to the crisis in reproducibility. Again, this is not intended to be an exhaustive treatment of the problem.

First, it is important to emphasize that these results do not represent fraud. Reports of fraudulent data are extremely rare in science. However, another benefit of enhancing reproducibility is that in a culture that emphasizes it, successful fraud will become nearly impossible to achieve.

We begin by considering technical contributions to the problem, then turn to the more important issue—the current culture and practice of science.

A) Publication bias ↥

In fields that rely on statistical analysis, such as biomedicine and psychology, the results published in refereed journals are not a representative sample of all results obtained through research because of a central feature of traditional, frequentist statistics. In frequentist statistics one can reject the null hypothesis (claim an effect exists) but one cannot accept the null hypothesis (claim an effect does not exist). As a result, the vast majority of publications report effects rather than non-effects. For example, despite the fact that the set of studies chosen for attempted replication in the Reproducibility Project: Psychology was NOT based on statistical significance in any way (see description above), *only three of the 100 original published studies* reported a null effect.

Errors will always occur when drawing conclusions from statistical analyses. A type I error occurs when a study produces a statistically significant effect when, in fact, that effect does not exist—in other words, it is a false positive. Almost always, Type I errors fail to replicate. A type II error occurs when one fails to detect an effect when it does exist—in other words, it is a false negative. Type II errors are much less likely to be reported than Type I errors even though the system is designed so that, in principle, Type II errors are more likely to occur. Clearly, the conservative strategy is to err on the side of not making a claim as opposed to making a false claim.

Unfortunately, publication bias greatly amplifies the proportion of studies reporting Type I errors. Simply to give a sense of the problem, we can consider a highly idealized situation. Assume for simplicity that all failures to reject the null hypothesis remain unreported and that the probability of a type I error is 5%. Now consider an effect that does not exist, but would be of great theoretical interest if it did exist. If 10 different researchers independently test for that effect in separate studies, there is about a 40% chance that one of them will obtain a significant result (i.e., $1.00-0.95^{10} = 0.401$). Thus, the likelihood that a false, non-reproducible result will be reported is **much** greater than 5%.

B) Small sample sizes ↥

The smaller the sample size, the more likely it is that the data do not give an accurate reflection of the true situation. In principle, as one collects more data, the probability of a Type I error goes to zero. The practical problem is that data can be expensive. The degree to which this is an issue varies widely across disciplines. In many areas of cognitive psychology, adding more data is virtually free (e.g., it is paid for in "research credits" for students). At the other extreme, in medical clinical trials, the cost of adding a patient can run into the thousands of dollars. As a result, sample sizes tend to be as small as is reasonably possible.

C) Cultural contributions to the reproducibility problem ↥

These proximal contributions to the reproducibility crisis are exacerbated by the current cultural climate in science. Arguably the most important cultural factor is the reward system. There is very strong pressure on scientists to produce "important" results. Obviously, "important" results are important, but they make up only a small subset of all the information that needs to be collected in an ideal strategy for advancing science. Currently, conducting replications gets almost no reinforcement from the reward system. Furthermore, information that is ultimately important, but whose importance is not immediately obvious, often has to be marketed as "important" support for some current theoretical position in order to be published. The alternative source of reward, quantity of output, also raises issues.

To make matters worse, this climate permeates every component of the system—funding, acceptance for publication, promotion and evaluation, and status in one's field—and operates in positive feedback loops among them. Funding agencies are under political pressure to be cost-effective in awarding grants, journals seek to increase their impact factors, and universities need some way to evaluate their employees (and naturally turn to the standards set by the various scientific communities). The relationship between the importance of scientists' findings and their status in their fields is obvious. We will discuss only a few of the ways this culture undermines reproducibility. We will also pass over its other deleterious effects.

1) Importance of results. A major contribution of culture to publication bias involves the interaction of the emphasis on "importance" and the human tendency to seek confirming evidence (Wasson, 1960). As noted earlier, the more often a non-existent effect is sought, the higher the rate of published Type I errors. The emphasis on important questions means some "would be important" effects that don't exist are almost certain to be "found" eventually. Such results then bolster confidence in the theoretical constructs they putatively support, regardless of the validity of those constructs. When researchers then seek to confirm invalid constructs which enjoy such support, they will inevitably investigate more "would be important" effects that don't exist.

2) Publication bias. Returning to publication bias, the technical proscription against accepting the null hypothesis, while an important contributor, is actually not the principle reason for the paucity of null effects in the literature. A number of statistical tools are available if one wishes to argue for no effect. Until very recently, journal space was a limited, precious commodity. As noted earlier, null effects receive little attention. Devoting limited journal space to rarely cited papers lowers a journal's impact factor. The advent of online journals solves the space problem in principle. However, online journals, both individually and as a group, need to

prove their worth by garnering high impact factors, which is not easily achieved by publishing null results.

The overarching pressure to find significant results in turn contributes to additional practices that lead to non-reproducibility

a) Sample size artifacts. When increasing sample size leads to an appreciable increase in costs (however defined), it is common to start with a small sample and to continue increasing sample size if needed until significance is achieved. This is called optional stopping and is the scientific equivalent of "best two out of three." As noted earlier, spurious results are more likely with a small sample size. In this practice, spurious results will be accepted before a more powerful sample size correctly rejects the effect. Optional stopping appears to inflate Type I error rates several-fold (Strube, 2006). This practice is part of a larger phenomenon called *p*-hacking (Simmons et al., 2011).

b) *P*-hacking. Flexibly selecting aspects of analysis based on what produces the most significant results is often referred to as *p*-hacking. It takes many forms, all of which involve *post hoc* analysis choices and all of which increase the likelihood of reporting an effect that does not exist—a false positive. We focus on a few possibilities to illustrate the idea. Data commonly requires "cleaning," that is to say removal of data values that are due to some extraneous source. This is often truly necessary to reveal patterns that are due to the process of interest. For example, EEG data contains artifacts due to eye blinks that must be removed (Luck, 2014). Unfortunately, subjective, *post hoc* choices are often necessary, opening the door to biases.

As another example, within a larger omnibus analysis, there are often pairwise comparisons that could be of interest. Sometimes a comparison stands out as potentially significant after data have been collected. Statistics provides corrections that allow such *post hoc* comparisons in a principled way. However, it is easy to reason "I would have planned it if I'd thought about it" and report the effect using the uncorrected statistic.

Finally, there often are a number of alternative analyses that could logically be applied to a particular set of data. In such cases, one should select analysis *a priori*. However, it is common practice to select the analysis after the data are known. The good news is that *p*-hacking is usually avoidable. To some extent it occurs because of a lack of statistical sophistication on the part of researchers.

D) Inadequate understanding of statistical probability and statistics training ↑

The American Statistical Association (ASA) recently issued a report on the current state of affairs in the use of statistical methods in science. They argue that many of the problems stem from a poor understanding of probability and statistics on the part of scientists. As noted above, many of the errors that underlie non-reproducibility are failures to follow principles that were taught, but perhaps not internalized, in graduate school. Furthermore, there are other useful statistical tools now widely available. For example, modern computational power now allows Bayesian null hypothesis testing. Hopefully a recent review of the ASA's report, published in Science, will bring these issues to the attention of the wider scientific community.

E) What would scientific Utopia look like? ↥

Nosek and colleagues (2012a, 2012b) argue that many of the problems facing science today, such as poor reproducibility, arise from using antiquated methods. They argue that careful investigations of the root causes of the problems, along with the potential of the ongoing revolution in Information and Communication Technologies (ICTs), and creative attempts at solutions, can help science come closer to its ideals—a Scientific Utopia if you like. The next section of the current report describes some of these proposed new standards. It focuses on those ideas that 1) have the strongest consensus support, and 2) have the best-developed infrastructure needed to begin implementation. Since most of the work to date has been undertaken by the Center for Open Science (COS) and its affiliates, we begin there.

IV) The Center for Open Science (COS) ↥

The Center for Open Science was founded in 2013 in Charlottesville, Virginia with the mission "to increase openness, integrity, and reproducibility of scientific research." It is supported by a number of organizations including: the Laura and John Arnold Foundation, the Alfred P. Sloan Foundation, the National Institutes of Health, and the National Science Foundation. We describe two of their most important services below in detail and briefly describe other services.

A) Transparency and Openness Promotion (TOP) guidelines ↥

The TOP guidelines program was described in *Science* by a committee brought together by the Center for Open Science. The committee included journal editors, representatives of funding agencies, and disciplinary leaders. Their goal was to address the problem of publication bias by setting guidelines to which journals could adhere. There are eight standards, each of which has three levels of implementation. Two standards reward researchers for the time and effort they have put into open practices; four standards specify what openness means across the scientific process; and, two standards address the preregistration process.

One example of these standards, described in some detail, helps to illustrate the nature of this enterprise. Qualified journals can issue badges to authors who meet specified openness standards in their published article. There are three such badges.

1) Open Data badge. This badge indicates that all data related to the published study are made available in an open-access repository (a list of over 2600 open-access repositories can be found at OpenDOAR). The data must be date-time stamped and in an immutable, permanent form, and must be accompanied by a "codebook" that allows others to work with the data. Finally, there must be an open license allowing others to use the data, while the licensor retains copyright. Full consideration is given to the need for legitimate exceptions, such as confidential information about human participants.

2) Open Materials badge. This badge indicates that all the information relevant to conducting an exact reproduction attempt is made available. This means that all digitally-shareable material must be made available subject to the same conditions as data. Non-shareable material and its relation to procedures must be described in sufficient detail for another

researcher to reproduce the study.

3) Preregistered and Preregistered+ badges. These badges allow readers to distinguish between *a priori* hypotheses and plans versus HARKing (Hypothesizing After Results are Known) and *post hoc* analyses. To earn these badges, researchers must, before any data are collected, submit a date- and time-stamped design which includes 1) a motivating question or hypothesis; 2) a description of all variables, including dependent variables, independent variables, controls and covariates; 3) planned sample size; and 4) a description of any materials to be used, or, when possible, the materials themselves. To earn the plus designation, a full planned analysis must also be included. It is important to note that preregistration does not preclude reporting of *post hoc* analyses, as long as they are reported as such.

There are currently over 500 journal signatories to the TOP guidelines, including journals published by: American Association for the Advancement of Science, Association for Psychological Science, BioMed Central, Public Library of Science, and The Royal Society.

There are about 60 organization signatories, including a few professional organizations such as the American Geophysical Union, the American Society for Cell Biology, and the Association for Psychological Science. There are also organizations devoted to open science that are pursuing new models, such as F1000Research, Mozilla's Science Lab, OpenFMRI, and Reddit's /r/openscience.

B) The Open Science Framework (OSF) ↑

The open practices described above make extensive use of open-access repositories and registries. A number of institutions provide these services, including: OSF, figshare, ClinicalTrials.gov, AEA registry, EGAP, ASCL, and others.

We focus on OSF because it is not domain specific and offers a number of services in addition to its study registry (which currently has over 4,000 entries). Most importantly, the projects service allows research teams to manage every aspect of a research project in a secure, cloud-based environment. Users control what parts are public and which are private. Users can also bring a variety of third party services (e.g., Dropbox, figshare, GitHub) together in one place. One especially nice feature is that users can create a registration, which is a permanent read-only, date-time stamped copy of the project being registered. This registration can be shared and even cited.

C) Dataset Crowdsourcing ↑

Dataset crowdsourcing is a new concept. Crowdsourcing is the method used to build and maintain Wikipedia. The idea in science would be to "recruit multiple independent analysts to investigate the same research question on the same data set in whatever manner they see as best... If everyone comes up with the same results, then scientists can speak with one voice. If not, the subjectivity and conditionality on analysis strategy is made transparent."

An initial project is investigate the question: "Are soccer referees more likely to give red cards to dark skin toned players than light skin toned players?" This project is being used to work out the most effective protocols for future crowdsourced projects.

D) The Collaborative Replications and Education Project (CREP) ↥

CREP is a crowdsourced replication project for undergraduates. It is intended to serve two purposes. First, it allows undergraduates to participate in research projects that may contribute to the advancement of science (and be published) while still putting primary emphasis on the educational value of the project for the student. Second, it could, if it becomes popular, become an important source of replication information. There are currently six projects underway.

E) Statistical and Methodological Consulting ↥

This service provides one-on-one e-mail and live Google Hangout consulting; and, online and on-site (on request) workshops. The University of Arizona recently hosted such a workshop. Services include *all* aspects of methodology and statistical analysis.

F) OSF for Meetings ↥

This is a free poster and presentation sharing service for use by academic meetings and conferences. It provides a page for each registered conference. The intention is to broaden awareness of upcoming meetings, streamline the application process, and increase the impact of meetings and conferences by making their presentations accessible online.

V) Journals ↥

Of all the institutions involved in producing science, journals have been the most proactive in terms of facilitating the practice of open science. To date, 538 journals have signed on to the TOP guidelines and a significant number have begun implementation of preregistration to varying degrees. Open science journals tend to be recently founded, online journals (e.g., PLOS ONE), but in some cases, established journals have started online versions that are aimed at publishing open science. Royal Society Open Science and Science Advances are notable examples.

VI) Professional Organizations ↥

In contrast, professional organizations have been slow to commit to promoting open science practices. Compared to journals, few professional organizations have signed on to the TOP Guidelines. The Association for Psychological Science is farthest along in promoting open science.

Professional organizations play a very critical role in promoting best science practices. Scientific method and culture vary widely across disciplines. There are few, if any, universals at the level of detailed implementation. Professional societies are acutely aware of the specifics of a particular discipline and are thus in the best position to determine what best scientific practices look like in their domain.

VII) Funding sources ↥

Some charitable organizations, for example the Laura and John Arnold Foundation, are financially supporting the move toward open science. Not surprisingly, tech companies committed to the FOSS ideal (Free Open Source Software) have been supportive of open science. Apache's Open Climate Workbench is one example.

However, the most important sources of funding are Federal granting agencies.

NIH has brought its efforts to promote rigor and reproducibility together in a dedicated area under Research & Training.

The effort to promote open science at NSF is not as well organized. However, NSF is involved in a move toward openness and transparency, having submitted a proposed framework for addressing issues of reproducibility to the Office of Management and Budget (OMB) in December 2014. NSF posted an interview with Brian Nosek, Director of the Center for Open Science, in their Discoveries section, and hosted an event focused on the issue of reproducibility in the summer of 2015.

Science is transnational. A Policy Paper from the Organization for Economic Cooperation and Development (OECD), published in October 2015, addresses the role of Government funding agencies world-wide in promoting open Science.

Finally, COS has proposed guidelines for funders.

VIII) What can universities do? ↥

A) Become a signatory of the Transparency and Openness Promotion guidelines

The TOP guidelines seem to be self-evidently desirable guidelines. Of course, any statement requires careful scrutiny before a university endorses it. We believe universities like ASU should give serious consideration to joining Carnegie Mellon as a signatory of the TOP guidelines. Contacts at CMU could presumably provide guidance regarding potential issues.

This could be more than just symbolic. The announcement of such an endorsement (along with an explanation of the issue and why the university supports efforts to promote better scientific practices) would go a long way toward raising awareness of the issue at the university.

B) Facilitate and promote awareness and training

A very common theme in discussions of the reproducibility crisis is the use of inappropriate statistical methods and other misuses of statistics. Resources exist for statistical consulting, yet many, if not most, researchers are either unaware of these resources or don't utilize them. Furthermore, many, if not most, researchers are not aware of the larger issue of best scientific practices and the many resources that have recently become available to facilitate a move toward better scientific practices.

Of course, many aspects of training and awareness must be handled at the level of specific units. Universities can help here by promoting awareness that this is becoming an important issue and that units need to address it in a manner that best fits their constituent disciplines. Of course, a university can also help by providing funds for unit-based training and awareness.

In addition, there are issues that are more universal and could be handled at the university (or college) level. For example, one big advantage of open science will be the growing availability of open access data repositories. A few basic skills can allow a researcher to get maximum benefit from this resource. A university that provides training in these skills could see a return on its investment in the form of increased research productivity.

The Center for Open Science offers workshops on a variety of relevant topics.

C) Assure adequate availability of open access repository storage

Storage costs money. Furthermore, on-site repositories can require extensive infrastructure. Fortunately, there are many open access repositories available off-site. OpenDOAR provides access to a very large number of repositories world-wide. Many repositories also provide open access for storage needs, generally at a price.

However, a top-tier research university must provide on-site open access archiving facilities. To be of full benefit, these facilities need to meet the standards specified in the TOP guidelines (e.g., date- and time-stamping), which are also the standards required by an increasing number of journals.

D) Participate in the promotion of best scientific practices

There is a reason journals have been the fastest to embrace the new ideals. An individual can practice science according to whatever standard he or she wants, and there will be a journal that will publish the work. Individuals can choose to publish in journals consistent with their standards. As new standards evolve, journals will evolve or appear in order to accommodate the new standards. However, universities must thoughtfully apply standards for valuing work when hiring and promoting researchers. An individual scientist can't just pick up and move if a university changes its standards for evaluating research. Because changes in evaluation policy can have a significant impact on the careers and reputations of faculty and academic professionals at a university, any changes must be made with great care. Many changes will produce winners and losers and thus face opposition no matter what the decision might be. Likewise, changes can cause unnecessary anxiety. Fields that are already closest to the ideal, such as some of the physical sciences, may worry that changes motivated by problems in other fields may impact them negatively. Of course, a wise policy would not do so. But it would also need to assuage those fears.

We believe that major changes are coming to many areas of science and that universities will play a critically important role in this change. However, it will not be the university's role to decide what those changes should be. It will be wisest to leave such deliberations to Professional organizations and to follow their lead on substance.

What universities can do now is to commit to taking seriously the process of change that is just beginning, and to begin thinking about innovative ways to respond.

The Reproducibility Subcommittee of the Research and Creative Activities Committee

Greg Stone (chair)
Rida Bazzi
Mary H. Burleson
Joseph Comfort
Hugh Mason
Arnold Maltz (*ex officio*)

References

Begley CG, Ellis LM. (2012). Raise standards for preclinical cancer research. *Nature* 483:531–533. doi: 10.1038/483531a.

Errington et al. (2014). An open investigation of the reproducibility of cancer biology research. *eLife*. DOI: 10.7554/eLife.04333.

Etz A, Vandekerckhove J. (2016). A Bayesian perspective on the Reproducibility Project: Psychology. *PLoS ONE* 11(2): e0149794. doi: 10.1371/journal.pone.0149794

Gorgolewski KJ; Wheeler K, Halchenko YO, Poline J-B, Poldrack RA. (2015). The impact of shared data in neuroimaging: The case of OpenfMRI.org. doi: 10.7490/f1000research.1110040.1

Ioannidis JPA. (2005). Why most published research findings are false. *PLOS Med*. 2, e124. doi: 10.1371/journal

Klein RA, et al. (2014). Investigating variation in replicability: A ‘‘Many Labs’’ replication project. *Social Psychology*, 45(3):142–152. doi: 10.1027/1864-9335/a00.

Luck SJ. (2014). An introduction to the event-related potential technique, 2nd Edition, Cambridge MA: MIT Press.

van Assen MALM, van Aert RCM, Nuijten MB, Wicherts JM. (2014). Why publishing everything is more effective than selective publishing of statistically significant results. http://dx.doi.org/10.1371/journal.pone.0084896

Nosek BA, Spies JR, & Motyl M. (2012). Scientific Utopia: II. Restructuring incentives and practices to promote truth over publishability. *Perspectives on Psychological Science*, 7(6) 615–631. doi: 10.1177/1745691612459058.

Nosek BA & Bar-Anan Y. (2012). Scientific Utopia: I. Opening Scientific Communication. *Psychological Inquiry*.

Nosek, et al. (2015). Estimating the reproducibility of psychological science. *Science*, 349: http://dx.doi.org/10.1126/science.aac4716

OECD. (2015). Making open science a reality. OECD Science, Technology and Industry Policy Papers, No. 25, OECD Publishing, Paris. http://dx.doi.org/10.1787/5jrs2f963zs1-en.

Prinz F, Schlange T, Asadullah K. (2011). Believe it or not: How much can we rely on published data on potential drug targets? *Nature Reviews Drug Discovery* 10:712. doi:10.1038/nrd3439-c1.

Simmons JP, Nelson LD, & Simonsohn U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*. http://doi.org/10.1177/0956797611417632

Strube, MJ. (2006). SNOOP: A program for demonstrating the consequences of premature and

repeated null hypothesis testing. *Behavior Research Methods*, 38(1), 24–27. http://doi.org/10.3758/BF03192746

Wason P. (1960). On the failure to eliminate hypotheses in a conceptual task. *Quarterly Journal of Experimental Psychology* 12(3): 129–140. doi:10.1080/17470216008416717

Wasserstein RL & Lazar NA. (2016): The ASA's statement on *p*-values: context, process, and purpose. *The American Statistician*. doi:10.1080/00031305.2016.1154108.