February 9, 2016

To: PCAST Members and Interested Colleagues

From: Lloyd Etheredge [1]

Re: **Comment: NSF's (Untrustworthy) Self-Correction Plan**

In 2015 the National Science Foundation published on its Website a scientific self-correction plan to address problems of unreliable research in economics and other social and behavioral sciences. I enclose a copy because the Report illustrates the untrustworthiness of NSF's senior leadership and scientific performance, **even when billions of people must suffer until more reliable economic science is available**. [2]

- The evidence ("Trusting the National Science Foundation will not solve this problem") mandates oversight review and swift corrective action, including a briefing to President Obama, at PCAST's level. Also, corrective steps by AAAS, journalists and Editors, and other system-level actors committed to reliable and trustworthy science.

Background

The Report was mandated by growing alarm and pressure about the unreliability of economic science since 2008, by suspicions about published psychological experiments, and by the broader scientific alarms (including NIH) raised by the research of Ioannidis and others. The original mandate was to "assess the scope and magnitude of the problem" of robustness (p. 2) and develop plans to correct these problems by NSF's Social, Behavioral, and Economic Sciences Directorate and others.

However, under the current NSF regime, this analysis was whittled-down and scientific standards were lowered: 1.) There was no professional review of published research; 2.) Notwithstanding earlier design specifications, data about the scope and alarming magnitude of the problems of unreliable macro-economic models – including data about problems known to NSF and uncorrected for many years - were excluded, along with all other data;[3] 3.) NSF limited itself to a closed, small, one-day workshop;

---

[1] Director, Government Learning Project, Policy Sciences Center, Inc., a public foundation. URL: www.policyscience.net; lloyd.etheredge@policyscience.net; 301-365-5241.

[2] The Report of the Subcommittee on Replicability of Science was prepared under the auspices of the Social, Behavioral and Economics (SBE) Advisory Committee. http://www.nsf.gov/sbe/AC_Materials/SBE_Robust_and_Reliable_Research_Report.pdf which reviewed the Report and authorized its public release.

[3] The Report speculates about innocent causes, but NSF has failed to correct known problems of unreliable economic science for several decades, even in the face of brutal confrontations by Committees of Visitors, formal charges to the NSF Inspector General about known missing variables, and professional feedback from government

4.) Economists were excluded from planning and participation in the study and workshop; 5.) The nation's social, behavioral, and economic scientists were not told about the study process. There were no public hearIngs and their input about the problems, magnitudes, and causes – that might have improved or challenged the Report – were not considered; 6.) There was no public comment period for drafts of national policy recommendations. (These were not made public until the Report was approved and published online.) Essentially, the Advisory Committee decided simply to rewrite a textbook, conceptual, discussion of the many possible sources of unreliable science, suppress known data about the scope and magnitude of the problem (i.e., including the awkward "known to NSF and uncorrected for many years" macro-economic problems) [4] and recommend a strategy of more studies in every direction.

NSF's methods created a distorted, unreliable analysis and failed to consider rapid and hopeful strategies to achieve robust and trustworthy macro-economic models. **NSF needed to do the competent literature review and evidence-based analysis because it is not difficult to improve the reliability of economic science.** (The paper that I sent to you earlier, "The Optimistic Case for Rapid Learning Economics," draws from the universe of published literature and lessons that NSF's *seigneurs* decided to omit – e.g., CBO time-series comparisons and lessons of GDP two-year forecasting errors of government and about 50 Blue Chip models since the late 1970s.)[5] **There is substantial professional agreement about several kinds of missing variables. We have good ideas about where to look for them.** There is a more powerful, informed, useful, competent, reliable, and hopeful Report about scientific reliability and rapid progress that the NSF system did not write.

Earlier, in discussing NSF's performance problems, I suggested to you the advice of the Boston Globe's Editor, portrayed in the movie Spotlight about untrustworthy priests. [I.e., Don't get into cat fights about blame in specific cases: who appointed an SBE scientific self-correction Committee without economists; who decided that national policy recommendations without competent literature reviews, and excluding data about the scope and magnitude of the problem, was an acceptable NSF standard;

users (e.g., CBO). The NSF-SBE Division and Advisory Committee that prepared the Report has conflicts of interest – i.e., a causal role in the uncorrected and growing unreliability of economic science since 2008.

[4] See also the view of John Ioannidis. In Benedict Carey, "Many Psychology Findings Are Not as Strong as Claimed, Study Says" The New York Times, August 27, 2015: Ioannidis is quoted as saying that the 50% non-confirmation problem reported for 100 psychological experiments "could be even worse in . . . economics."

[5] A reference copy of the paper is online at www.policyscience.net.

who “trusted their colleagues” or “trusted their subordinates” too much; who steered the agenda to exonerate the NSF system from a causal analysis of its miserable performance since the late 1970s, etc.] Instead, I think it is time for PCAST to recognize and solve the institutional and system-level problems.[6]

To underscore, again, the urgency: **Billions of people will continue to be injured until the problem of reliable economic science is solved**. The request to the NSF Director and to the National Science Board for a self-correction plan for the Social, Behavioral, and Economic Sciences was a test of their scientific integrity, competence, and professional trustworthiness. They failed. Many accountable (and morally obtuse) people should be replaced, beginning at the top. Then, with other system-level changes, we can have a trustworthy NSF and a more hopeful future.

Attachment

NSF-SBE, Robust and Reliable Research Report (2015).

[6] System-level solutions include, for example, the Code of Journalistic Ethics for Science that removes duress and creates expectations for a watchdog role in the relationship to NSF.

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

Subcommittee Members:

Kenneth Bollen (University of North Carolina at Chapel Hill, Cochair)
John T. Cacioppo (University of Chicago, Cochair)
Robert M. Kaplan (Agency for Healthcare Research and Quality)
Jon A. Krosnick (Stanford University)
James L. Olds (George Mason University)[1]

Staff Assistance

Heather Dean (National Science Foundation)

May, 2015

[1] James Olds came off the Subcommittee in October, 2014 when he became the Assistant Director for the Directorate for Biological Sciences at NSF.

INTRODUCTION

Scientific knowledge is cumulative. The production of each empirical finding should be viewed more as a promissory note than a final conclusion. If a scientific finding cannot be independently verified, then it cannot be regarded as an empirical fact. And if a literature contains illusory evidence rather than real findings, the efficiency of the scientific process can be compromised.

In recent years, we have seen an accumulation of evidence suggesting that some scientific findings thought to be robust may in fact be illusory (e.g., Ioannidis, 2008). In some instances, initial findings turned out to be intentionally fraudulent, maliciously fabricated rather than being generated through genuine data collection and analysis. Scientists presume that such outright fraud is rare, because instances of it have seldom emerged.

But with the passage of time, an increasing number of studies suggest that conventional scientific practices, including practices routinely taught to students learning to become scientists, may sometimes yield findings that are not reliable because they are the result of well-intentioned data collection, data management, or data analysis procedures that unintentionally lead to conclusions that are not robust.

Although the behaviors that yield illusory findings appear to be occurring across many scientific disciplines, an understanding of these behaviors and the development of measures that can prevent them seem especially well-suited to social, behavioral, and economic scientists. Social and behavioral scientists routinely study the causes of human behaviors and the effectiveness of strategies meant to change behavior. Furthermore, the National Science Foundation (NSF) Directorate for Social, Behavioral and Economic Sciences (SBE) is positioned to establish policies and fund research to mitigate the factors that affect the robustness of scientific research.

In the spring of 2013, the NSF SBE Advisory Committee (AC) established a subcommittee to investigate actions NSF SBE might pursue to promote robust research practices in science. This document constitutes the subcommittee's report.

BACKGROUND

During the summer and fall of 2014, the subcommittee designed a proposal for a workshop on "Robust Research in the Social, Behavioral, and Economic Sciences." This workshop was convened on February 20-21, 2014, at the National Science Foundation, the objectives of which were to: (a) assess the scope and magnitude of the problem, and review and critique the extant recommendations and solutions to promote scientific replicability; (b) foster a reflective and extended dialog among researchers, journal editors, and science administrators about what integrated set of policies and procedures might be acceptable and beneficial to the scientific community; (c) identify a set of recommendations to optimize the incentives for laudatory scientific behavior while minimizing unintended side effects; and (d) position SBE to support research exploring the causes and consequences of scientific behaviors that enhance the likelihood of generating nonreplicable findings and replicable findings, and into research practices to improve the validity of research findings.

A variety of individuals and academic bodies (e.g., funding agencies, scientific associations, journals) have done work on this topic, and this work was considered when constituting the membership of and goals for the workshop. The participants in the workshop were drawn from multiple disciplines and were asked to consider: (i) the scope and magnitude of the problem of nonrobust scientific practices in science, (ii) possible improvements in scientific practice and procedures, (iii) the implications for science education and training, (iv) the implications for editorial policies and procedures, (v) the implications for research university policies and evaluation criteria, and (vi) the implications for federal funding policies and evaluation criteria.

The attendees included experts on each of these issues, but our goal for the workshop was not simply to re-discuss well-known perspectives that had been disseminated during prior conferences or in existing publications. Instead, we asked presenters to: (a) review *briefly* what they saw to be the problems and possible solutions; (b) address the possible costs and unintended side-effects of possible solutions, including the differential costs or impacts on investigators who are engaged in robust scientific practices versus those who may be more susceptible to the situational pressures that impact replicability; and (c) offer recommendations about research that NSF could fund to improve the replicability, validity, generativity, and integration of research across all sciences.

The charge to the subcommittee and the workshop agenda and summary are provided in the Appendices. Our purpose here is to extract some of the key issues that emerged and to outline recommendations for a research agenda that might improve the robustness of research across the sciences. The report is organized as follows. The next section defines key terms. This is followed by several recommendations on reproducibility, replicability, and generalizability. Additional recommendations concern issues of statistical power, confirmation bias, and understanding scientific research as it is practiced. The conclusions follow the recommendations.

DEFINITIONS

We view *robust* scientific findings as ones that are *reproducible*, *replicable*, and *generalizable*. Reproducibility, replication, and generalizability are different though related concepts that are vital to our discussion. In practice, these ideas are sometimes confounded or combined. Because writers do not always use these terms in the same way, we explain our usage. We make no claim to be providing the true meaning of these concepts, but do hope that these definitions clarify our meanings of the terms.

Reproducibility refers to the ability of a researcher to duplicate the results of a prior study using the same materials and procedures as were used by the original investigator. So in an attempt to reproduce a published statistical analysis, a second researcher might use the same raw data to build the same analysis files and implement the same statistical analysis to determine whether they yield the same results. For example, a study might involve OLS regressions conducted using data from the 2014 American National Election Study survey. After publication of the results, another investigator using the same data can attempt to conduct the same analyses. If the same results were obtained, the first set of results would be deemed reproducible. If the same

results were not obtained, the discrepancy could be due to differences in processing of the data, differences in the application of statistical tools, differences in the operations performed by the statistical tools, accidental errors by an investigator, and other factors. Reproducibility is a minimum necessary condition for a finding to be believable and informative.

Replicability refers to the ability of a researcher to duplicate the results of a prior study if the same procedures are followed but new data are collected. That is, a failure to replicate a scientific finding is commonly thought to occur when one study documents relations between two or more variables and a subsequent attempt to implement the same operations fails to yield the same relations with the new data. It also is possible and useful to demonstrate the lack of relations between variables. For instance, an intervention might fail to produce an effect. A second study might investigate the same intervention in a different setting and also find no effect. Thus, null results can also be replicated.

When a single researcher is conducting his or her own study for a second time, it might seem easy to repeat the same data collection and analysis procedures, because the researcher is fully informed about the procedures. But when another researcher in another location did not observe the conduct of the first study and relies on a textual description of its data collection and analysis procedures, critical details may not be fully and effectively understood, so the procedures implemented second may not match the procedures implemented in the initial study. Thus, an apparent failure to replicate a finding may occur because importantly different procedures are used the second time.

More generally, failure to replicate can occur for a number of reasons, including: (1) the first study's methods were flawed, and the variables are not related to one another, (2) the second study's methods were flawed; the variables are truly related to one another, but this was misleadingly not revealed by the second study, (3) the two studies do not disagree with one another, because the association observed in the first study is not statistically significantly different from the association observed in the second study, once we take into account the sampling fluctuations that occur in both studies, or (4) the methods or participants used in the second study are substantively different from those used in the first study, so the second does not match the first in terms of key conditions (e.g., different types of people participated in the first and second studies).

Generalizability refers to whether the results of a study apply in other contexts or populations that differ from the original one. Generalization can be done from one set of human participants in an experiment to other people (e.g., findings generated with college student participants at a large Midwest university might be generalized to the entire U.S. adult population). Generalization can also be done from one persuasive message that was studied in a laboratory experiment to all persuasive messages that could be presented to people in the course of daily life. Generalizability concerns the degree to which found relations apply in different situations. Usually a finding's failure to generalize indicates the operation of limiting conditions, the identification of which advances theory.

RECOMMENDATIONS

With the definitions in hand, we make recommendations to address the robustness of scientific research. The first few recommendations are classified under our concepts of reproducibility, replicability, and generalizability. The others focus on statistical power, confirmation bias, and understanding the scientific research in practice.

Reproducibility

Science should routinely evaluate the reproducibility of findings that enjoy a prominent role in the published literature. To make reproduction possible, efficient, and informative, researchers should sufficiently document the details of the procedures used to collect data, to convert observations into analyzable data, and to analyze data. Therefore, our first recommendation is:

> *Recommendation 1: Each report of research supported by NSF should be accompanied by detailed documentation on procedures to enable an independent researcher to reproduce the results of the original researcher. A report of what these archives contain and how they are accessible should be required in a project's Final Report and in descriptions of "Prior NSF Funding" in proposals seeking new support.*

Ideally, all materials used to collect data, to transform data, and to analyze data would be archived in a public accessible online storage facility. Records of all statistical analysis code and all statistical analysis output should be included in the archive. Any materials collected using paper or other tangible methods should be stored or electronically recorded (e.g., photographs), and any procedures should be videotaped when implemented. These electronic archives should also be made publicly available via the Internet. If materials are purchased, purchase sources and purchase specifications (e.g., for measuring devices) should be recorded in the archive.

If issues of confidentiality preclude sharing all raw data, perhaps summary statistics can be generated (e.g., a matrix of correlations among measured variables) to allow other researchers to reproduce findings using such statistics.

Replicability

If a researcher attempts to replicate a study by using similar procedures to collect new data and similar analytic tools, the similarity of the findings to those of the original study can be compared. Although we have an intuitive sense of what it means for results to replicate, the meaning becomes less clear the more closely we look. One way to judge replication would be that the results are identical across studies. That is, the effect of a manipulation on an outcome variable should be of the same size and significance. Or the correlation between two variables should be of the same size and significance. However, this is quite a strict approach and likely unrealistic.

Another approach would be to calculate a confidence interval around the estimates generated by the two studies and assess whether the confidence intervals overlap. Though this has some

intuitive appeal, there are problems with such an approach (Schenker & Gentleman, 2001). Another possibility is to estimate the same association in the two studies under the constraint that both are equal and to compare the fit of a model that allows each effect to differ in the two different studies. An even more relaxed approach would be to require that an association be of the same sign and statistical significance in the two studies to conclude that replication occurred. Yet another approach would be to focus on effect sizes or other standardized measures of associations between variables and to define replication as obtaining similar effect sizes.

In light of these ambiguities, we offer the following recommendations:

> *Recommendation 2: NSF should sponsor research that evaluates various approaches to determining whether a finding replicates and to assess which approach(es) under which circumstances are the most helpful for reaching valid conclusions about replicability.*
>
> *Recommendation 3: To permit assessing replication in various ways, NSF should encourage researchers to report associations between variables using different metrics (e.g., standardized and unstandardized coefficients, effect sizes, odds ratios) and indicating precision of estimates (with standard errors) and to assess the statistical significance of findings using these different methods.*

Generalizability

All research occurs in a context that has at least some unique conditions. It could be the population from which the sample is drawn. It could be a particular combination of variables that enhances or depresses effects. In some situations, one variable might substitute for another in bringing about an effect. Alternatively, two or more antecedent conditions might be required to produce an outcome. Thus, the size of an effect can differ across studies when: (a) other (initially unidentified) antecedents vary across these studies, or (b) one or more moderator variables are operating across these studies. Consequently, effect sizes can vary depending on the experimental control over, or contributions of, other influences for the outcome of interest. For instance, a genetic marker may show a strong association to a phenotype when all other factors are held constant, whereas the same genetic marker may show a very weak association to the phenotype in genome-wide association studies when multiple genetic, epigenetic, situational, and gene x environmental interactions influence the phenotype. Inconsistent findings might therefore result from a failure to generalize across the dissimilar conditions in diverse studies rather than a lack of relationships between variables.

Another example comes from research on attitude change (and laboratory research on social behavior more generally). Findings appeared not to replicate, because the same experimental factors (e.g., source credibility) were found to produce different outcomes in different studies. Rather than treat this as a statistical or methodological problem, two distinct mechanisms (routes) were identified through which attitude change could occur, and the theoretical conditions were specified in which a given factor or set of factors would trigger each route. The resulting Elaboration Likelihood Model (Petty & Cacioppo, 1981, 1986) made sense of what had

appeared to be conflicting results and generated predictions of new patterns of data that were subsequently verified. Thus, lack of generalization led to theoretical maturation.

Careful attention to study details (conceptualization, operationalization, experimental control over other potential independent variables, statistical power, execution, analysis, interpretation) increases the likelihood that empirical results constitute robust scientific facts upon which one can build. Minimal robustness suggests that an empirical effect has been established, and failures to replicate the finding using different measures, situations, time points, or populations suggest the operation of potentially important moderator variables (and, thus, generate theoretical questions). Failure to find the same results across these facets of a research design may reflect a failure to generalize and may trigger a search for the operation of a previously unrecognized determinant or moderator variable. Treating such discrepancies as raising theoretical questions rather than simply noting that studies differed in terms of methodology should foster the development of testable hypotheses and, ultimately, more comprehensive theories.

One specific type of difference between studies in the social, behavioral, and economic sciences involves the participants in the research. Some studies are done with college students enrolled in psychology courses at large universities, other studies are done with college students enrolled at small, elite colleges, other studies are done on websites at which people volunteer to participate in research for little or no compensation, and other studies make use of probability samples of precisely specified populations.

Some disciplines have historically treated differences between participant samples as nuisance variables, presumably unrelated to the findings of a study, which are presumed to generalize across all people. Other disciplines have historically placed great value on representative random samples and placed little faith in the generalizability of the findings of studies of haphazard samples of rare subpopulations. Some disciplines have assumed that the findings of laboratory studies are generalizable to real-world, uncontrolled settings outside the lab, whereas other disciplines have believed that insights into real-world thinking and action must be gained by studying cognition and behavior in its natural settings (e.g., studying voting in real national elections instead of in constructed lab settings). Lastly, some disciplines presume that findings transcend time, so results obtained today should be obtained in a replication attempt ten years from now. In contrast, other disciplines place more significance on the impact of temporal context on findings and would treat a replication attempt years after a first study as an attempt to generalize a finding across time. As a result, when inconsistent results are observed across studies, different disciplines reach different conclusions about the likely causes, some calling these instances of lack of replication, while others call them instances of lack of generalizability.

To facilitate gaining insight in the face of such puzzles, we recommend:

> *Recommendation 4: NSF should sponsor research that identifies optimal procedures for practically assessing all types of generalizability of findings (e.g., from a set of study participants to a population, from one set of measures to other measures, from one set of circumstances to other circumstances) and differentiating lack of generalizability from failure to replicate.*

Statistical Power

Consider again the case in which a study is conducted twice, the results appear to be different (e.g., one study yields a statistically significant treatment effect and the other does not), but a test of the two effects suggests that there is no statistically significant difference. One might conclude that the results replicate. However, in truth, this pattern of data may be illusory, the result of insufficient statistical power. The smaller is the number of observations analyzed, the larger is the standard error of the estimated effect. So the uncertainty resulting from small samples might reduce the statistical power to detect differences in effects even when they are present.

Small samples can also cause problems in another way. An experiment can be run with a small sample relatively quickly and easily. So if the results of a small sample run of an experiment are not what an experimenter expects to see, it is minimally costly to discard the data on the grounds that "something must have gone wrong in the implementation" and conduct the experiment again. Multiple runs of an experiment increase the chance that an apparently statistically significant finding will appear which is in fact an illusory result of chance-alone variation in results across experiments.

Stated more generally, studies with small samples and minimal statistical power are likely to yield inaccurate pictures of reality when combined with only a subset of these findings being reported (e.g., Button et al., 2013). Studies with small samples reduce the probability of detecting a true effect (due to low statistical power), increase the probability that the effect size of a true effect is overestimated (due to the use of $p < .05$ to identify when an effect has been "detected" and the larger sampling error associated with smaller sample sizes), and increase the probability that an apparently statistically significant effect is not truly different from zero (due to differences in the base rates for tests of true and untrue effects). Because initial effect size influences calculations of the needed statistical power for replications, replication attempts with ample statistical power to detect a *reported* (i.e., over-estimated) effect may be underpowered to detect the *true* effect.

One means of increasing statistical power is to increase sample size. Increasing sample size while holding all other variables constant increases the precision of an effect size estimate (i.e., statistical power) by decreasing standard errors. For many researchers, however, increasing sample size may be very difficult and costly, as when studying rare subpopulations. And generally stated, collecting larger samples of data requires larger research budgets.

One illustration of this problem is in the arena of studies using functional magnetic resonance imaging (fMRI). These studies tend to involve very few participants, due to the high cost of collecting data from each participant. Therefore, these studies are routinely underpowered. In order for fMRI studies to be well powered, the total budget supporting such work would need to be substantially increased, or the number of such studies would need to be decreased.

One might imagine that quantitative meta-analyses of such studies can yield enhanced statistical power by combining data from large numbers of participants. However, authors' reporting habits inhibit the effectiveness of such meta-analyses. In order to be included in a meta-analysis,

a study's report must provide exact estimates of effect sizes and p-values or other such statistics. Unfortunately, however, use of the arbitrary cut-off of $p<.05$ has often led researchers to report simply whether a p-value is above or below that threshold and not to report sizes of effects that are not statistically significant. Consequently, meta-analyses of fMRI studies can aggregate only the effect sizes for studies in which the test of an effect reached statistical significance. Given the small sample size in most neuroimaging research, small but theoretically important effects are therefore likely to go undetected (due to low statistical power), thereby providing at best an incomplete and at worst a misleading depiction of underlying neural mechanisms. A solution to this problem may be relatively simple: creating and enforcing standards for full reporting of the results of statistical analyses to allow comprehensive and precise meta-analyses.

> *Recommendation 5: NSF should fund research exploring the optimal and minimum standards for reporting statistical results so as to permit useful meta-analyses.*

Confirmation Bias

Confirmation bias refers to a tendency to search for or interpret information in a way that confirms one's preconceptions or hypotheses, to avoid exposure to challenging information, and to discredit the challenging information one does encounter. Much research has documented confirmation bias in people's acquisition and processing of information.

In that light, it should come as no surprise that scientists may also manifest confirmation bias. Scientists may actively seek out and assign more weight to evidence that confirms their hypotheses and ignore or underweight evidence that could disconfirm their hypotheses. When the results of a study are not as expected, an investigator may be highly motivated to check over the data processing in search of accidental errors that can be corrected, whereas when expected results are obtained, such thorough scrutiny may be less likely, and errors may go undetected.

The computation of statistics when analyzing empirical data is not always governed by rules that clearly and specifically prescribe just one way to analyze a set of data. Most often, multiple different analytic approaches could be considered legitimate for a single application. For example, data might be legitimately analyzed using ordinary least squares regression or logistic regression. A continuous variable might be entered in its original metric in a regression, or it could be subjected to a log or square root transformation. These and other choices of researchers can produce sets of results that differ importantly in their implications.

Checking robustness of findings to seemingly arbitrary analytic approach differences is a recommended component of any investigation. But generating many different sets of results and selecting one to report simply because it confirms a researcher's expectations is a behavior referred to as "p-hacking": a disingenuous attempt to generate a publishable result when the full array of available evidence raises questions about its replicability.

A variety of other "questionable research practices" have been identified, including:

> (a) failing to report analyses of all of the measures collected in a study and describing

only those that yield desired findings;

(b) deciding whether to collect more data after determining whether obtained results with a smaller sample document desired results;

(c) failing to report analyses of data from all relevant experimental conditions that were executed in the course of data collection, because data from those conditions did not yield desired results;

(d) stopping collecting data earlier than initially planned because desired results have already been obtained;

(e) "rounding off" a *p* value in a way inconsistent with conventional practice (e.g., reporting that a *p* value of .054 is less than .05) in order to enhance the apparent robustness of a desired finding;

(f) reporting only studies that produced desired findings and discarding studies that did not produce desired findings;

(g) deciding to exclude data points only after determining that doing so will enhance the degree to which a study seems to produce desired findings;

(h) keeping in data points because without them the desired findings will no longer be found;

(i) reporting an unexpected finding as if it had been predicted *a priori* and thereby increasing its apparent plausibility;

(j) claiming that analytic results are unaltered by controlling for other variables when this has not been fully checked empirically.

The desire to produce findings in line with one's prior publications and to avoid discrediting one's own prior work, to earn tenure and promotion in academic settings, to be awarded grant funds, to gain visibility in scientific circles and beyond, and other forces may encourage researchers to engage in p-hacking, thus filling the literature with false findings. In general terms, undesirable researcher behaviors have been referred to as "questionable research practices": practices that can yield illusory findings though that can also be used to assure that findings are robust.

In light of the inefficiencies and inaccuracies that result from p-hacking and other forms of confirmation bias, we recommend:

Recommendation 6: NSF should support research into the use of questionable research practices, the causes that encourage such behavior, and the effectiveness of proposed interventions intended to discourage such behavior and should support the identification

of empirically-validated optimal research practices to avoid the production of illusory findings.

Recommendation 7: In NSF grant proposals, investigators should be required to describe plans for implementing and fully reporting tests of the robustness of findings using alternate analytical methods (when appropriate). In addition, researchers should be encouraged to design studies whose outcomes would be theoretically interesting regardless of the outcome, or of seriously considering more than one hypothesis. In grant progress reports and final reports, investigators should be required to describe whether more than one hypothesis was considered, the robustness checks conducted and results obtained.

Understanding Scientific Practice

In some fields, conventional practices have been widely adopted by investigators and have unintentionally caused findings to be illusory (e.g., Vul et al., 2009) or incorrectly interpreted (e.g., Jussim, 2012). Therefore, in addition to studying individual findings and their robustness, there is value in studying the research practices of various scientific disciplines, to explore whether any traditional practices might undermine the efficiency of theory development.

Consider, for example, an experiment to be conducted with members of a specific population. Participants might be randomly sampled from the population, they might be randomly assigned to either a treatment or control condition, the number of participants might be sufficient to yield the needed statistical power to detect an effect of the treatment, and optimal measures of outcome variables might be administered. But in practice, participants might not comply with the treatment regimen (e.g., taking an aspirin every single day), data may not be collected from all participants because some drop out of the study entirely or fail to provide needed assessments, distributions of data may violate the assumptions underlying the statistics computed, and tests of statistical significance may not properly take into account all sources of non-independence and uncertainty in observed patterns. In some fields such departures from the ideal are recognized and addressed directly by investigators, while in other fields such departures are largely overlooked.

SBE scientists are especially well equipped to understand the development and maintenance of field-wide norms of conduct that undermine research robustness. In-depth interviews, participant observation, and other qualitative methods hold promise in deepening our understanding of scientific research in practice. Survey methods, administrative data, meta-analysis, and other quantitative research approaches permit us to tackle this issue from other angles. Working together, SBE scientists can document departures from the ideals of scientific investigation in practice that are widespread within fields. This in turn can provide insight into the major sources of non-robust research findings.

Recommendation 8: NSF should sponsor research seeking to document suboptimal practices that are widespread in particular fields, with an eye towards identifying those

areas that most depart from the scientific ideals and contribute to nonrobust research findings.

CONCLUSION

Science is a cumulative process that hinges on the repeated investigation of the same questions from many angles. Because each study builds on the insights produced by prior studies, efficiency of the scientific process can be substantially compromised if a literature includes studies that report illusory results, produced either intentionally or unintentionally. Similarly, scientific efficiency will be compromised if failures to produce expected findings are misdiagnosed as failures to reproduce earlier results or failures to replicate earlier studies when in fact they are failures of earlier findings to generalize.

It is therefore in the interest of all sciences to:

1) Identify questionable research practices that cause illusory findings to make their way into the published literature.

2) Encourage attempts to reproduce, replicate, and generalize scientific findings.

3) Conduct careful investigations when attempts to reproduce, replicate, and generalize scientific findings fail, in order to correctly diagnose the causes.

4) Identify forces that encourage scientists to implement questionable research practices.

5) Propose and test interventions that are meant to reduce the frequency with which questionable research practices are implemented.

Much can be done by NSF SBE to promote all of the above, and we look forward to it doing so. In light of upcoming work we hope will be conducted, we recommend:

Recommendation 9: NSF should create a Foundation-wide committee of experts to monitor issues of reproducibility, replicability, and generalizability of findings, to support investigations of these issues and disseminate insights gained both within the Foundation and outside the Foundation, to propose ways to change the NSF granting process to enhance scientific quality and efficiency, and to provide leadership on these issues in the coming decades.

References

Bacon, E (1960). *The Novum Organum and Related Writings.* New York: Liberal Arts Press. (Original work published 1620)

Brazier, M. A. (1959). The historical development of neurophysiology. In J. Field (Ed.), *Handbook of physiology. Section I: Neurophysiology.* (Vol. 1, pp. 1-58). Washington, DC: American Physiological Society.

Button, K., Ioannidis, J., Mokrysz, C., Nosek, B., Flint, J., Robinson, E., & Munafò, M. (2013). Power failure: why small sample size undermines the reliability of neuroscience. *Nature reviews: Neuroscience, 14*(5), 365–76. doi:10.1038/nrn3475

Cacioppo, J. T., & Tassinary, L. G. (1990). Inferring psychological significance from physiological signals. *American Psychologist, 45,* 16-28.

Chandler, J., Mueller, P., & Paolacci, G. (2014). Nonnaïveté among Amazon Mechanical Turk workers: Consequences and solutions for behavioral researchers. *Behavior Research Methods, 46*(1), 112-130. doi: 10.3758/s13428-013-0365-7

Epley, N., & Dunning, D. (2001). Feeling "holier than thou": Are self-serving assessments produced by errors in self- or social prediction? *Journal of Personality and Social Psychology, 79*(6), 861-875. doi:10.10337//0022-3514.79.6.861

Epley, N., & Dunning, D. (2006). The mixed blessings of self-knowledge in behavioral prediction: Enhanced discrimination but exacerbated bias. *Personality and Social Psychology Bulletin, 32*(5), 641-655. doi: 10.1177/0146167205284007

Goodman, J., Cryder, C., & Cheema, A. (2013). Data collection in a flat world: The strengths and weaknesses of Mechanical Turk samples. *Journal of Behavioral Decision Making, 26*(3), 213224. doi:10.1002/bdm.1753

Ioannidis, J.P.A. (2005). Why most published research findings are false. *PLoS Medicine, 2*(8), e124. doi:10.1371/journal.pmed.0020124

Ioannidis, J.P.A. (2008). Why most discovered true associations are inflated. *Epidemiology, 19*(5), 640–648. doi:10.1097/EDE.0b013e31818131e7

John, L., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological science, 23*(5), 524–32. doi:10.1177/0956797611430953

Jussim, L. J. (2012). *Social Perception and Social Reality: Why Accuracy Dominates Bias and Self-Fulfilling Prophecy*. New York, NY: Oxford University Press.

Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B. Jr., Bahnik, S., et al. (2014). *Social Psychology, 45*, 142-152.

Orne, M.T. (1961). On the social psychology of the psychological experiment: With particular reference to demand characteristics and their implications. *American Psychologist, 17*(11), 776-783. doi:10.1037/h0043424

Petty, R. E., & Cacioppo, J. T. (1981). *Attitudes and Persuasion: Classic and Contemporary Approaches*. Dubuque, Iowa: Wm. C. Brown. (Reprinted 1996, Westview Press, Boulder, CO.)

Petty, R. E., & Cacioppo, J. T. (1986). The elaboration likelihood model of persuasion. *Advances in Experimental Social Psychology, 19*, 123-205.

Platt, J. R. (1964). Strong inference. *Science, 146,* 347-353.

Rand, D.G. (2011). The promise of Mechanical Turk: How online labor markets can help theorists run behavioral experiments. *Journal of Theoretical Biology, 299*, 172-179. doi:10.1016/j.jtbi.2011.03.004

Rosenthal, R. (1963). On the social psychology of the psychological experiment: The experimenter's hypothesis as unintended determinant of experimental results. *American Scientist, 51*(2), 268-283.

Schenker, N. & Gentleman, J. (2001). On judging the significance of differences by examining the overlap between confidence intervals. *The American Statistician*, 55, April 2001, 182-186.

Steering Committee of the Physicians' Health Study Research Group. (1988). Final report on the aspirin component of the ongoing Physicians' Health Study. *New England Journal of Medicine, 321*, 129-135.

Tourangeau, R. & Yan, T. (2007). Sensitive questions in surveys. *Psychologial Bulletin, 133*(5), 859-883. doi:10.1037/0033-2909.133.5.859.

Vul, Ed., Harris, C., Winkielman, P., & Pashler, H. (2009). Puzzlingly high correlations in fMRI studies of emotion, personality, and social cognition. *Perspectives on Psychological Science, 4*, 274-290.

Appendices

A. Subcommittee on Replicability in Science, Advisory Committee to the National Science Foundation Directorate for Social, Behavioral, and Economic Sciences: Charge and Members

B. *Robust Research in the Social, Behavioral, and Economic Sciences*: Workshop agenda

C. *Robust Research in the Social, Behavioral, and Economic Sciences*: Workshop summary

Appendix A

National Science Foundation (NSF)
Directorate for the Social, Behavioral, and Economic Sciences (SBE)
Replicability in Science
A Subcommittee of the Advisory Committee (AC) for the Social, Behavioral and Economic Sciences

Charge

The Subcommittee on Replicability in Science will work with NSF Staff and various communities to deliver recommendations to the SBE AC in Fall, 2013.

The scope of the subcommittee will include the following:

- Examination of the current state of knowledge regarding issues of replicability in the SBE sciences such as:
    - Institutional norms, including publication bias
    - Generalizability versus replicability
    - Research on robust research practices
- Identifying partners
- Consideration of the resources, both human and financial, needed to encourage robust research practices and replication of scientific findings
- Consideration of the relationship between the challenge of replicability and the potential offered by the recent Office of Science Technology Policy memo on enhanced access to data and publications
- Recommendations for future actions

Input to be used by the Subcommittee will include, but not be limited to:

- Data available from NSF and other sources;
- Input from individual members of the community, either obtained individually or through workshops and other forums;
- Discussions with program officers, other NSF staff, and NSF leadership across the Foundation.

Recommendations in the report should include, but not be limited to:

- Areas of science where future investment in replicability are likely to produce significant transformative increases in the likelihood that published data are replicable and trustworthy;

- Appropriate mechanisms for supporting replicability in the future. Examples would include individual research grants or funds for graduate training in statistics;

- Opportunities for collaboration across and beyond NSF to ensure increased likelihood of replicability in SBE sciences;

- Other comments the subcommittee deems relevant to the charge.

While the subcommittee is not required to announce its meetings and hold them open to the public because it will be reporting directly to the SBE AC, any documents received by or created by the subcommittee may be subject to access by the public. The subcommittee should deliver its final report and present a summary of this report for consideration of acceptance by the SBE AC at its meeting in November, 2013.

April 1, 2013

Subcommittee on Replicability in Science:
Kenneth A. Bollen, University of North Carolina, Chapel Hill (CoChair)
John T. Cacioppo, University of Chicago (CoChair)
Jon A. Krosnick, Stanford University
Robert M. Kaplan, Agency for Healthcare Research and Quality
James L. Olds, George Mason University
NSF SBE Liaison: Heather Dean

Appendix B

Robust Research in the Social, Behavioral, and Economic Sciences
Directorate for Social, Behavioral, and Economic Sciences (SBE)
Advisory Committee (AC) Subcommittee on Replicability in Science
National Science Foundation (NSF)

A Two-Day Workshop at NSF
February 20-21, 2014
Agenda

Day 1

8:15 am - 8:30 am: Welcome and Opening Remarks
Joanne Tornow, Acting Assistant Director, SBE
John Cacioppo, University of Chicago, SBE AC Subcommittee CoChair

8:30 am - 10:30 am: Panel I – Scope and Magnitude of the Problem and Recommendations for Scientific Practice: I
Moderator: Kenneth Bollen, University of North Carolina, Chapel Hill
Participants:
1. Brian Nosek, University of Virginia
2. Hal Pashler, University of California, San Diego
3. Patricia Devine, University of Wisconsin
4. Leslie K. John, Harvard University Business School

10:30 am - 12:30 pm: Panel II – Scope and Magnitude of the Problem and Recommendations for Scientific Practice: II
Moderator: James Olds, George Mason University
Participants:
1. Gregory Francis, Purdue University
2. David Funder, University of California, Riverside
3. Ron Thisted, University of Chicago
4. Katherine Button, University of Bristol

12:30 pm - 1:30 pm: Lunch Break

1:30 pm - 3:30 pm: Panel III – Education & Training
Moderator: Robert Kaplan, Office of Behavioral and Social Sciences Research, National Institutes of Health (OBSSR, NIH)
Participants:
1. Jo Handelsman, Yale University
2. Simine Vazire, Washington University in St. Louis
3. Richard Ball, Haverford College
4. Larry Hedges, Northwestern University

3:30 pm - 3:45 pm: Break

3:45 pm - 5:45 pm: Panel IV – Editorial/Journal Policies and Procedures
Moderator: Jon Krosnick, Stanford University
Participants:
1. Marcia McNutt, American Association for the Advancement of Science, *Science*
2. Bobbie Spellman, University of Virginia, *Perspectives on Psychological Science*
3. Eric Eich, University of British Columbia, *Psychological Science*
4. Giorgio Ascoli, George Mason University, *Neuroinformatics*

5:45 pm - 6:00 pm: Day 1 Wrap-up
John Cacioppo, University of Chicago

Day 2

8:30 am - 9:00 am: Review of Proposals with Pros & Cons
Moderator: Robert Kaplan, OBSSR, NIH

9:00 am - 11:00 am: Panel V – Institutional Policies and Procedures
Moderator: Kenneth Bollen, University of North Carolina, Chapel Hill
Participants:
1. Alan Kraut, Association for Psychological Science
2. Richard Saller, Stanford University
3. Gary VandenBos, American Psychological Association (APA)
4. Barbara Entwisle, University of North Carolina, Chapel Hill
5. Brad Hesse, National Cancer Institute, NIH; Chair, APA Publications Board

11:00 am - 1:00 pm: Lunch break

1:00 pm - 3:00 pm: Panel VI – Funding Agency Opportunities and Policies
Moderator: Jon Krosnick, Stanford University
Participants:
1. Elena Koustova, National Institute on Drug Abuse, NIH
2. Robert Kaplan, OBSSR, NIH
3. Richard Nakamura, Center for Scientific Review, NIH
4. Philip Rubin, Office of Science and Technology Policy

3:00 pm - 3:30 pm: Break

3:30 pm - 4:30 pm: Reflection and Group Discussion
Conclusions and Next Steps
Participant: Anthony Greenwald, University of Washington

Moderators & Discussion Leaders: John Cacioppo, University of Chicago, & James Olds, George Mason University

Appendix C

Robust Research in the Social, Behavioral and Economic Sciences
Directorate for Social, Behavioral, and Economic Sciences (SBE)
Advisory Committee (AC) Subcommittee on Replicability in Science
National Science Foundation (NSF)

A Two-Day Workshop at NSF
February 20-21, 2014
Workshop Summary

Welcome and Opening Remarks

Joanne Tornow, Acting Assistant Director, SBE, and John Cacioppo, CoChair of the SBE Advisory Committee Subcommittee on Replicability in Science, welcomed the participants, thanked the workshop organizers and described the goals of the workshop. They also articulated the importance of scientific replicability, not just to the SBE sciences, but to the entire scientific and engineering enterprise.

Panels I and II. Scope and Magnitude of the Problem and Recommendations for Scientific Practice

Panel I participants:
Brian Nosek, University of Virginia
Hal Pashler, University of California, San Diego
Patricia Devine, University of Wisconsin
Leslie K. John, Harvard University Business School
Kenneth Bollen, University of North Carolina, Chapel Hill (moderator)

Panel II participants:
Gregory Francis, Purdue University
David Funder, University of California, Riverside
Ron Thisted, University of Chicago
Katherine Button, University of Bristol
James Olds, George Mason University (moderator)

The first two panels of the workshop were devoted to identifying contributors to irreproducibility, the scope and magnitude of the problem, and possible solutions. The speakers reinforced the importance of replicability to the progress of science and to its credibility, not only for scientists, but for the taxpaying public and policy-makers as well. It was also noted that while scientific fraud receives a great deal of media attention, it is in fact extremely rare, and not the focus of the workshop's discussions.

With that as context, the panel presentations and discussions surfaced a range of other factors and scientific practices that contribute to irreproducibility:

- A scientific culture that incentivizes the publication of novel, positive results, relegating negative results and replication studies to the "file drawer";
- Questionable research practices such as
    - terminating studies as soon as the desired results are attained;
    - dropping observations, measures, items or conditions after looking at outcomes of interest;
    - running multiple experiments with similar procedures and reporting only those yielding significant results;
- Inadequate statistical power and sample sizes;
- Researcher bias;
- Subtle changes in methodology and execution when investigators attempt to repeat previously published studies; and
- Differences in study subject selection or setting when studies are repeated.

There was also considerable discussion of the terms, *replicability* and *reproducibility*. There is no consensus on their definitions and they are often used interchangeably to depict a variety of outcomes. These include duplicating a result when re-analyzing the original data from a study or when repeating an entire experiment with as much fidelity to the original as possible. In contrast, failure to obtain the same result when an experiment is repeated in a different setting or with different subjects may be an issue of *generalizability*, revealing important information about the phenomenon under investigation and prompting additional scientific exploration.

Another overarching consideration was how we conceptualize replication. Exact replication should not be expected. Similar experiments might be considered point estimates in distributions of results of repeated executions of the same experiment. It was also suggested that we move away from our "pass-fail" model of individual scientific studies and instead, consider studies in the context of a whole program of research. Statistical tools to estimate the probability that a study has produced a reliable result in the context of other studies on the same topic can facilitate this approach.

The speakers recommended numerous solutions to improve scientific replicability, several of which centered on increasing the transparency of the scientific process. This more open science would encompass:

- Sharing data, analysis code, and study materials;
- Providing methodological details of research;
- Disclosing all data exclusions, manipulations, and measures in studies; and
- Registering studies, requiring pre-specification of primary and secondary outcomes, study design, expected sample size, and data analysis plan.

A second set of solutions focused on data analysis and reporting in grant proposals and publications. Specific suggestions included the following:

- De-emphasizing statistical significance as an outcome or an objective /criterion for publication;
- Using Bayesian methods to account for multiple sources of variation;
- Requiring federal funding applicants to describe their inference populations, sampling methods, and methods to assess the match between sample and population;
- Contracting out statistical analysis of datasets to avoid conflicts of interest;
- Requiring manuscript authors to discuss sample sizes and statistical power, and report effect sizes and 95% confidence intervals for their studies; and
- Encouraging more meta-analysis as a formal process to quantify accumulating knowledge, and making studies "meta-analyzable" through use of standardized protocols, instruments and measures.

A third set of solutions targeted incentives and opportunities for replication and open science, such as:

- Providing funding for replication studies;
- Providing publication outlets for replication studies;
- Creating replication consortia;
- Incentivizing open science through the use of "badges" by journals to signal to readers that the study was registered, and that the authors had signed statements certifying the legitimacy of their execution and analysis of study procedures;
- Striking a balance between federal funding of groundbreaking work and definitive research, as the latter would more likely include replication; and
- Encouraging collaboration among investigators to increase statistical power and replicate findings.

Some of the recommended solutions are already being implemented. The Center for Open Science in Charlottesville, VA was founded in 2013. It builds and distributes tools and provides products and services to improve the openness, integrity, and reproducibility of scientific research. The Center's primary infrastructure is the Open Science Framework, which provides project management support to research teams through the entire research lifecycle: planning, execution, reporting, archiving and discovery. Another resource is Psych FileDrawer, a web archive of replication attempts in experimental psychology. The website is designed to make it quick and convenient to upload reports but also to require enough detail to make the report credible and responsible. The site also provides a discussion forum for each posting, allowing users to discuss the report (potentially allowing collective brainstorming about possible moderator variables, defects in the original study or in the non-replication attempt, etc.).

Scientific societies are also addressing irreproducibility. The Society for Personality and Social Psychology (SPSP) Task Force has recommended a number of changes in SPSP journal policies: having authors discuss sample size and statistical power; report effect sizes and 95% confidence intervals of findings; and include in an appendix the verbatim wording of instructions, manipulations and measures, in all manuscripts submitted for publication. In addition, the Task Force recommended development of a data sharing policy, explicit mention of replications as among the types of articles that journals will consider, and the creation of additional outlets for publication of replication studies.

Panel III: Education & Training

Panel participants:
Jo Handelsman, Yale University
Simine Vazire, Washington University in St. Louis
Richard Ball, Haverford College
Larry Hedges, Northwestern University
Robert Kaplan, Office of Behavioral and Social Sciences Research, National Institutes of Health (NIH; moderator)

Following the identification of the scope of the problem and potential solutions, the discussion turned to the need for better education and training to improve scientific replicability. The speakers in the third panel of the workshop offered numerous suggestions to do so:

- Improved training in ethics (with periodic refresher courses);
- Earlier, stronger and more consistent training across the SBE sciences in statistics, including instruction in effect sizes and confidence intervals, statistical power, and meta-analysis/meta-analytic thinking;
- Encouraging a culture of "getting it right" rather than "finding significant results" during training;
- Teaching transparency in data reporting, including the reporting of "imperfect" results, and telling the "whole story", rather than a "good story";
- Improving methodological education by teaching students to avoid QRPs;
- Educating students in replication and data-sharing and supporting the development of educational materials to do so;
- Cross-training in different fields to improve scientists' abilities to identify uncontrolled variables; and
- Using published data to teach students the scientific method, methodological transparency, and the importance of replication.

It was acknowledged that the gate-keepers of science (e.g., journal editors, university administrators, hiring and promotion committees, and funding agencies) would need to model, endorse and reward good scientific practices. There was also additional discussion of the magnitude of change needed to produce the desired result of more robust science: a wholesale restructuring of training in the SBE sciences vs. more modest changes in the curriculum.

Panel IV. Editorial/Journal Policies and Procedures

Panel participants:
Marcia McNutt, American Association for the Advancement of Science, *Science*
Bobbie Spellman, University of Virginia, *Perspectives on Psychological Science*
Eric Eich, University of British Columbia, *Psychological Science*
Giorgio Ascoli, George Mason University, *Neuroinformatics*
Jon Krosnick, Stanford University (moderator)

The fourth panel of the workshop presented the perspective of scientific journal editors. They confirmed that numerous scientific disciplines are grappling with the challenge of ensuring scientific replicability and described a variety of steps they are taking to address the challenge. These efforts take different forms:

- *Neuroinformatics* publishes original articles and reviews with an emphasis on data structure and software tools related to analysis, modeling, integration, and sharing in all areas of neuroscience research. In addition to the traditional original scientific articles, *Neuroinformatics* publishes "data original" articles. These are full length manuscripts reflecting an original, significant data contribution to the neuroscience field, that are fully referenced and abstracted, and peer reviewed.

- *Science* has adopted recommendations from the National Institute of Neurological Disorders and Stroke (National Institutes of Health) for the reporting of preclinical studies (Landis et al., 2012). At a minimum studies should report on sample-size estimation, whether and how animals were randomized, whether investigators were blind to the treatment, and the handling of data. In addition, the journal has added additional members to its Board of Reviewing Editors from the statistics community, and is hosting a series of workshops on replicability in different scientific disciplines.

- *Perspectives on Psychological Science* is starting to publish new types of articles: "Ideas to Watch", short papers describing ideas that are good and suggestive though not yet complete, and "Say it Ain't So", a series that will correct long-standing literature after it has moved forward. It has also initiated a Registered Replication Report article type that is a collection of independently conducted, direct replications of an original study, all of which follow a shared, predetermined protocol.

- *Psychological Science* has introduced five new initiatives aimed at raising the bar on publication standards and practices: removal of word limits on Methods and Results sections; clarification of criteria for manuscript evaluation; a badge system to promote open scientific practices (e.g., open data, open materials, or pre-registered studies); placing less emphasis on null hypothesis significance testing and more on effect sizes, confidence intervals, and meta-analysis; and enhancing the reporting of research methods. *Psychological Science* also conducted a disclosure statement pilot experiment, the aims of which were to assess authors' willingness to disclose methodological information that is not normally reported under current publication guidelines, and to develop a clear picture of what disclosure statements would look like, should the journal decide to require them in the future. The project's results suggested that disclosure statements could deliver important information about research methodology (e.g., data exclusions, dropped manipulations, or dropped measures, and how sample size was determined) and can be completed quickly, without significantly impacting manuscript submission rates.

Panel V. Institutional Policies and Procedures

Panel participants:

Alan Kraut, Association for Psychological Science
Richard Saller, Stanford University
Gary VandenBos, American Psychological Association (APA)
Barbara Entwisle, University of North Carolina, Chapel Hill
Brad Hesse, National Cancer Institute, NIH; Chair, APA Publications Board
Kenneth Bollen, University of North Carolina, Chapel Hill (moderator)

The workshop's fifth panel presented the perspectives of university administrators and professional scientific societies. The university representatives reinforced observations made earlier in the workshop, i.e., that research misconduct (fraud) is extremely rare, and that irreproducibility is not unique to the SBE sciences. Participants also noted that the incentive structures of universities may be contributing to scientific irreproducibility, including the perceptions of junior researchers about what is valued by their institutions, and expectations for tenure. Their recommendations for universities' roles in ensuring robust science include the following:

- Reinforcing the value of research integrity in their faculty appointment and promotion processes;
- Encouraging best practices through mentoring younger faculty, postdocs, and graduate students; and
- Developing models for data storage, sharing and archiving.

The representatives from the Association for Psychological Science (APS) and American Psychological Association (APA) provided additional examples of activities that demonstrated that the SBE sciences are at the forefront of discussions about scientific replicability:

- The APS special issue on replicability and good research practices has been downloaded 500,000 times, and by researchers well beyond psychological science;
- The APA Task Force on Replication in the Psychological Literature is developing guidelines to specify criteria for a "good" replication study;
- APA has authorized its journals to have independent, online-only, peer-reviewed "replication sections", and established the *Archives of Scientific Psychology* as an open methods, collaborative, data-sharing, open access journal.

Speakers in this panel articulated additional benefits of data sharing. In addition to enabling replications, data sharing promotes aggregation for knowledge synthesis, hypothesis generation, programmatic decisions, and generalizability testing. It provides opportunities for data analysis with more powerful analytic techniques than were available when the data were originally collected. They also identified a number of additional issues to explore in discussions of scientific robustness:

- Ensuring the protection of human subjects in the context of data sharing and data repositories;
- Supporting the evolution of the scientific publishing enterprise through development and deployment of infrastructure, including the development of easy-to-use tools for researchers, and services designed to improve replication and data sharing; and

- Re-examining an incentive structure that encourages investigators to create new and innovative instruments that results in the collection of data that are difficult to harmonize.

Panel VI. Funding Agency Opportunities and Policies

Panel participants:
Elena Koustova, National Institute on Drug Abuse, NIH
Robert Kaplan, Office of Behavioral and Social Sciences Research, NIH
Richard Nakamura, Center for Scientific Review, NIH
Philip Rubin, Office of Science and Technology Policy
Jon Krosnick, Stanford University (moderator)

The sixth panel in the workshop was devoted to federal agencies' roles in enhancing scientific reproducibility. The NIH Office of Behavioral and Social Sciences Research (OBSSR) is responsible for promoting and coordinating behavioral and social sciences research across the agency and has been working on the problem of replication since 2011. In early 2012, OBSSR, in collaboration with the National Institute on Aging, National Institute of Mental Health, National Library of Medicine and APS held a meeting of thought leaders to discuss the topic. This effort spawned the Registered Replication Report initiative in *Perspectives on Psychological Science* that was described earlier in the workshop. OBSSR is also working with an international group to develop CONSORT-SPI, an extension of the CONSORT Guidelines for social and psychological interventions. CONSORT (Consolidated Standards of Reporting Trials) encompasses various initiatives to alleviate the problems arising from inadequate reporting of randomized, controlled, clinical trials.

In addition, the NIH leadership published a paper in Nature, outlining its initiatives to enhance reproducibility of pre-clinical research (Collins and Tabak, 2014). These efforts are focused on raising community awareness; enhancing formal training; improving the evaluation of grant applications; protecting the integrity of science by adoption of more systematic review processes; and increasing stability for investigators. Specific NIH activities to address these needs include the following:

- Experimenting with checklists to ensure more systematic evaluation of experimental design and analysis in grant applications;
- Piloting assignment of specific reviewers to evaluate the scientific premises on which grant applications are based;
- Developing a new training module on enhancing reproducibility and transparency for intramural postdoctoral fellows and for broader dissemination;
- Launching PubMed Commons, a pilot program testing options for scientists to post online comments on original research articles; and
- Considering grant mechanisms that allow more flexibility and a longer period of support than currently available, to provide greater stability for investigators at certain career stages.

The discussions during this panel also revealed concerns about some of the recommended solutions to improve replicability. The use of simple checklists for peer review of research grant

applications' methodologies, for example, might have the unintended consequence of stifling scientific creativity. An overzealous shift toward replications studies could slow the progress of science. Participants also reiterated the importance of engaging the scientific community in any development of solutions to improve replicability.

Conclusions and Next Steps

Participant: Anthony Greenwald, University of Washington
John Cacioppo, University of Chicago, and James Olds, George Mason University (moderators)

The final session of the workshop was a lively discussion of many of the earlier ideas for making research more robust and replicable. It also produced a number of additional recommendations:

- Establish a code of research ethics in the *process* of research;
- Publish papers contingent on successful replications; and
- Use new techniques to detect questionable research practices.

These discussions also produced suggestions for a research agenda that NSF might pursue to improve scientific replicability:

- Methodological research to improve scientific replicability;
- Research to validate candidate best and questionable research practices;
- Replications of important research findings; and
- Empirical research to test the effectiveness of the intervention strategies to improve replicability.

The workshop concluded with the thanking of all of the participants and an outline of the next steps. First, the Subcommittee on Replicability will report on the workshop at the spring, 2014 meeting of the SBE Advisory Committee (AC). After consideration of the discussions and recommendations from the workshop, the Subcommittee will address its original charge in a full report to the SBE AC at its fall, 2014 meeting. Once the report is finalized and formally accepted by the AC, it will be made public on the NSF website.

References

Landis S.C., Amara, S.G., Asadullah, K., Austin, C.P., Blumenstein, R., Bradley, E.W., Crystal, R.G., Darnell, R.B., Ferrante, R.J., Fillit, H., Finkelstein, R., Fisher, M., Gendelman, H.E., Golub, R.M., Goudreau, J.L., Gross, R.A., Gubitz, A.K., Hesterlee, S.E., Howells, D.W., Huguenard, J., Kelner, K., Koroshetz, W., Krainc, D., Lazic, S.E., Levine, M.S., Macleod, M.R., McCall, J.M., Moxley III, R.T., Narasimhan, K., Noble, L.J., Perrin, S., Porter, J.D., Steward, O., Unger, E., Utz, U., & Silberberg, S.D. (2012). A call for transparent reporting to optimize the predictive value of preclinical research. *Nature, 490*, 187-191.

Collins F.C. & Tabak L.A. (2014). NIH plans to enhance reproducibility, *Nature, 505*, 612-613.