Statistical Challenges in Assessing and Fostering the Reproducibility of Scientific Results: Summary of a Workshop

DETAILS

132 pages | 7 x 10 | PAPERBACK
ISBN 978-0-309-39202-0 | DOI: 10.17226/21915

AUTHORS

Michelle Schwalbe, Rapporteur; Committee on Applied and Theoretical Statistics; Board on Mathematical Sciences and Their Applications; Division on Engineering and Physical Sciences; National Academies of Sciences, Engineering, and Medicine

Statistical Challenges in Assessing and Fostering the Reproducibility of Scientific Results

Summary of a Workshop

Michelle Schwalbe, *Rapporteur*

Committee on Applied and Theoretical Statistics

Board on Mathematical Sciences and Their Applications

Division on Engineering and Physical Sciences

The National Academies of
SCIENCES • ENGINEERING • MEDICINE

THE NATIONAL ACADEMIES PRESS
Washington, DC
www.nap.edu

THE NATIONAL ACADEMIES PRESS 500 Fifth Street, NW Washington, DC 20001

This workshop was supported by Grant No. DMS-1351163 between the National Academies of Sciences and the National Science Foundation. Any opinions, findings, or conclusions expressed in this publication do not necessarily reflect the views of any organization or agency that provided support for the project.

International Standard Book Number-13: 978-0-309-39202-0
International Standard Book Number-10: 0-309-39202-0
Digital Object Identifier: 10.17226/21915

Cover: The photo "Peas in a Pod" by Andrew_Writer is used with Creative Commons Attribution 2.0 Generic license (https://www.flickr.com/photos/dragontomato/3698617520/in/gallery-106928713@N04-72157638971870524/).

This report is available in limited quantities from:

Board on Mathematical Sciences and Their Applications
National Academies of Sciences, Engineering, and Medicine
500 Fifth Street NW
Washington, DC 20001
bmsa@nas.edu
http://www.nas.edu/bmsa

Additional copies of this workshop summary are available for sale from the National Academies Press, 500 Fifth Street, NW, Keck 360, Washington, DC 20001; (800) 624-6242 or (202) 334-3313; http://www.nap.edu/.

Printed in the United States of America

Suggested citation: National Academies of Sciences, Engineering, and Medicine. 2016. *Statistical Challenges in Assessing and Fostering the Reproducibility of Scientific Results: Summary of a Workshop.* Washington, DC: The National Academies Press. doi: 10.17226/21915.

The National Academies of
SCIENCES • ENGINEERING • MEDICINE

The **National Academy of Sciences** was established in 1863 by an Act of Congress, signed by President Lincoln, as a private, nongovernmental institution to advise the nation on issues related to science and technology. Members are elected by their peers for outstanding contributions to research. Dr. Ralph J. Cicerone is president.

The **National Academy of Engineering** was established in 1964 under the charter of the National Academy of Sciences to bring the practices of engineering to advising the nation. Members are elected by their peers for extraordinary contributions to engineering. Dr. C. D. Mote, Jr., is president.

The **National Academy of Medicine** (formerly the Institute of Medicine) was established in 1970 under the charter of the National Academy of Sciences to advise the nation on medical and health issues. Members are elected by their peers for distinguished contributions to medicine and health. Dr. Victor J. Dzau is president.

The three Academies work together as the **National Academies of Sciences, Engineering, and Medicine** to provide independent, objective analysis and advice to the nation and conduct other activities to solve complex problems and inform public policy decisions. The Academies also encourage education and research, recognize outstanding contributions to knowledge, and increase public understanding in matters of science, engineering, and medicine.

Learn more about the National Academies of Sciences, Engineering, and Medicine at **www.national-academies.org**.

PLANNING COMMITTEE ON
STATISTICAL CHALLENGES IN ASSESSING AND FOSTERING
THE REPRODUCIBILITY OF SCIENTIFIC RESULTS

CONSTANTINE GATSONIS, Brown University, *Co-Chair*
GIOVANNI PARMIGIANI, Dana-Farber Cancer Institute, *Co-Chair*
STEPHEN FIENBERG, Carnegie Mellon University
STEVEN N. GOODMAN, Stanford University School of Medicine
JOHN H. HOLMES, University of Pennsylvania Perelman School of Medicine
ALAN F. KARR, RTI International
JELENA KOVAČEVIĆ, Carnegie Mellon University
XIHONG LIN, Harvard University
ROGER PENG, Johns Hopkins Bloomberg School of Public Health
VICTORIA STODDEN, University of Illinois, Urbana-Champaign

Staff

MICHELLE K. SCHWALBE, Senior Program Officer
RODNEY N. HOWARD, Administrative Assistant
LINDA CASOLA, Senior Program Assistant and Staff Editor

Acknowledgment of Reviewers

This workshop summary has been reviewed in draft form by individuals chosen for their diverse perspectives and technical expertise, in accordance with procedures approved by the Report Review Committee. The purpose of this independent review is to provide candid and critical comments that will assist the institution in making its published workshop summary as sound as possible and to ensure that the workshop summary meets institutional standards for objectivity, evidence, and responsiveness to the project's charge. The review comments and draft manuscript remain confidential to protect the integrity of the study process. We wish to thank the following individuals for their review of this workshop summary:

C. Glenn Begley, TetraLogic Pharmaceuticals,
Joel Greenhouse, Carnegie Mellon University,
Joelle Lomax, Science Exchange, and
Randy Schekman, University of California, Berkeley.

Although the reviewers listed above have provided many constructive comments and suggestions, they were not asked to endorse the views presented at the workshop, nor did they see the final draft of the workshop summary before its release. The review of this workshop summary was overseen by Lawrence D. Brown, University of Pennsylvania, who was responsible for making certain that an independent examination of this workshop summary was carried out in accordance with institutional procedures and that all review comments were carefully considered. Responsibility for the final content of this summary rests entirely with the author and the institution.

Contents

1

Introduction

Questions about the reproducibility of scientific research have been raised in numerous settings and have gained visibility through several high-profile journal and popular press articles. Quantitative issues contributing to reproducibility challenges have been considered (including improper data management and analysis, inadequate statistical expertise, and incomplete data, among others), but there is no clear consensus on how best to approach or to minimize these problems.

This is an issue across all scientific domains. A recent study found that 65 percent of medical studies were inconsistent when retested, and only 6 percent were completely reproducible (Prinz et al., 2011). The following year, a survey published in *Nature* found that 47 out of 53 medical research papers on the subject of cancer were irreproducible (Begley and Ellis, 2012). The Begley and Ellis *Nature* study was itself reproduced in the journal *PLOS ONE*, which confirmed that a majority of cancer researchers surveyed had been unable to reproduce a result.

A lack of reproducibility of scientific results has created some distrust in scientific findings among the general public, scientists, funding agencies, and industries. For example, the pharmaceutical and biotechnology industries depend on the validity of published findings from academic investigators prior to initiating programs to develop new diagnostic and therapeutic agents that benefit cancer patients. But that validity has come into question recently as investigators from companies have noted poor reproducibility of published results from academic laboratories, which limits the ability to transfer findings from the laboratory to the clinic (Mobley et al., 2013).

While studies fail for a variety of reasons, many factors contribute to the lack of perfect reproducibility, including insufficient training in experimental design, misaligned incentives for publication and the implications for university tenure, intentional manipulation, poor data management and analysis, and inadequate instances of statistical inference. The workshop summarized in this report was designed not to address the social and experimental challenges but instead to focus on the latter issues of improper data management and analysis, inadequate statistical expertise, incomplete data, and difficulties applying sound statistical inference to the available data.

As part of its core support of the Committee on Applied and Theoretical Statistics (CATS), the National Science Foundation (NSF) Division of Mathematical Sciences requested that CATS hold a workshop on a topic of particular importance to the mathematical and statistical community. CATS selected the topic of statistical challenges in assessing and fostering the reproducibility of scientific results.

WORKSHOP OVERVIEW

On February 26-27, 2015, the National Academies of Sciences, Engineering, and Medicine convened a workshop of experts from diverse communities to examine this topic. Many efforts have emerged over recent years to draw attention to and improve reproducibility of scientific work. This workshop uniquely focused on the statistical perspective of three issues: the extent of reproducibility, the causes of reproducibility failures, and the potential remedies for these failures. CATS established a planning committee (see p. v) to identify specific workshop topics, invite speakers, and plan the agenda. A complete statement of task is shown in Box 1.1.

The workshop, sponsored by NSF, was held at the National Academy of Sciences building in Washington, D.C. Approximately 75 people, including speakers, members of the planning committee and CATS, invited guests, and members of the public, participated in the 2-day workshop. The workshop was also webcast live to nearly 300 online participants.

This report has been prepared by the workshop rapporteur as a factual summary of what occurred at the workshop. The planning committee's role was limited to organizing and convening the workshop. The views contained in the report are those of individual workshop participants and do not necessarily represent the views of all workshop participants, the planning committee, or the National Academies of Sciences, Engineering, and Medicine.

In addition to the summary provided here, materials related to the workshop can be found online at the website of the Board on Mathematical Sciences and Their Applications (http://www.nas.edu/bmsa), including the agenda, speaker presentations, archived webcasts of the presentations and discussions, and other background materials.

BOX 1.1
Statement of Task

An NRC [National Research Council]-appointed program committee will plan and organize a workshop to address statistical challenges in assessing and fostering the reproducibility of scientific results. To this end, the workshop will examine three issues from a statistical perspective: the extent of reproducibility, the causes of reproducibility failures, and potential remedies. Specifically,

- What are appropriate metrics and study designs that can be used to quantify reproducibility of scientific results?
  —Variability across studies is a well-known phenomenon and has given rise to the field of research synthesis and meta-analysis. How should this variability be assessed? What degree of variability would lead to concerns about lack of reproducibility?

- How can the choice of statistical methods for study design and analysis affect the reproducibility of a scientific result?
  —How does routine statistical hypothesis testing with widely used thresholds for test significance affect the reproducibility of results? How do standard methods for study design and choice of sample size affect reproducibility?

- Are there analytical and infrastructural approaches that can enhance reproducibility, within disciplines and overall?
  —Do we need new conceptual/theoretical frameworks for assessing the strength of evidence from a study? Do we need broad adoption of practices for making study protocols and study data available to the scientific community? How can this be achieved?

In addressing these three issues, the workshop will

- Bring together representatives from different disciplines,
- Illustrate case studies, and
- Include some participants who are in positions to affect the incentive systems in the right direction.

One or more rapporteurs who are not members of the committee will be appointed to create a workshop summary report.

WORKSHOP THEMES

Over the course of the workshop, speakers discussed possible reasons as to why studies may lack reproducibility. The following topics were discussed repeatedly throughout the workshop: clarifying definitions of reproducibility and associated terms, improving scientific discovery, increasing the accepted threshold for statistical significance, enhancing and clarifying protocols, uniting the broad scientific community in reproducibility efforts, changing research incentives, increasing

sharing of research material, and enhancing education and training. The discussions around each of these areas are summarized in this section.

Clarifying Terminology

Throughout the workshop, presenters (Yoav Benjamini, Ronald Boisvert, Steven Goodman, Xiaoming Huo, Randy LeVeque, Giovanni Parmigiani, Victoria Stodden, and Justin Wolfers) and participants referenced the confusion in the terminology associated with reproducibility. Below are some of the terms and definitions that were offered:

- *Reproducibility*. "The ability of a researcher to duplicate the results of a prior study using the same materials . . . as were used by the original investigator. . . . A second researcher might use the same raw data to build the same analysis files and implement the same statistical analysis . . . [in an attempt to] yield the same results. . . . 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." (NSF, 2015 [as identified by Steven Goodman])
- *Repeatability* (also referred to as *empirical reproducibility*). The ability to see the data, run the code, and follow the specified steps, protocols, and designs as described in a publication. (Steven Goodman and Victoria Stodden)
- *Replicability*. "The ability of a researcher to duplicate the results of a prior study if the same procedures are followed but new data are collected. . . . A failure to replicate a scientific finding is commonly thought to occur when one study documents [statistically significant] relations between two or more variables and a subsequent attempt to implement the same operations fails to yield the same [statistically significant] relations." (NSF, 2015 [as identified by Steven Goodman])
- *Robustness*. The resistance of the quantitative findings or qualitative conclusions to (minor or moderate) changes in the experimental or analytic procedures and assumptions. (Steven Goodman)
- *Statistical reproducibility*. The notion of how statistics and statistical methods contribute to the likelihood that a scientific result is reproducible and to the study and measurement of reproducibility. (Victoria Stodden)
- *Computational reproducibility*. Any issue arising from having a computer involved somewhere in the work process, from researchers who do bench work and analyze their data with a spreadsheet, to researchers doing work

on large computing systems with an enormous amount of code and software. (Victoria Stodden)

Improving Scientific Discovery

Several speakers discussed the importance of enhancing reproducibility to improve scientific discovery (Micah Altman, Steven Goodman, Randy LeVeque, Giovanni Parmigiani, and Marc Suchard). In order to improve discovery, evidence must be generated to help the scientific community reach consensus about a question of interest. While it is rare that a single study will provide sufficient evidence to yield a consensus, a key step in this process is replication—generating evidence under different experimental settings and across different populations. The accumulation and weighing of such evidence informs the process by which the scientific community reaches consensus about the question of interest. A central component of this process includes the systematic elimination of alternative explanations for observed associations and the explicit acknowledgement that as new evidence arises, the consensus in the scientific community might change.

Presenter Steven Goodman discussed two additional advantages of strengthening replication: (1) increased understanding of the robustness of results, including their resistance to (minor or moderate) changes in the experimental or analytic procedures and assumptions; and (2) increased understanding of the generalizability (i.e., transportability) of the results, including the truth of the findings outside the experimental frame or in a not-yet-tested situation. Goodman added that the border between robustness and generalizability is indistinct because all scientific findings must have some degree of generalizability.

Increasing the Threshold for Scientific Significance

Several speakers (Dennis Boos, Andreas Buja, Steven Goodman, Valen Johnson, and Victoria Stodden) and participants discussed the inadequacy of the current p-value standard of 0.05 to demonstrate scientific significance. Some alternative proposals included reducing the standard p-value (Buja) by at least an order of magnitude (Johnson), switching to a p-value range (Boos), or switching to a Bayes factor equivalent (Johnson). There was some opposition to changing the standard, specifically due to the possibility that additional resources would be needed to meet the requirement for larger sample sizes.

Enhancing and Clarifying Protocols

Multiple speakers (Micah Altman, Andreas Buja, Marcia McNutt, Joelle Lomax, Mark Suchard, and Victoria Stodden) discussed the importance of protocols through-

out the workshop, including experimental methodology, data analytical actions (e.g., model selection and tuning), and coding decisions (e.g., instrumentation design and software).

Uniting the Community in Reproducibility Efforts

The need for a unified, multifaceted approach for dealing with reproducibility was emphasized by multiple speakers (Chaitan Baru, Philip Bourne, Steven Goodman, Mark Liberman, Marcia McNutt, Victoria Stodden, Irene Qualters, and Lawrence Tabak). They argued that this effort must include all stakeholders, including funding agencies, journals, universities, industries, and researchers.

Changing Research Incentives

Community incentives for reproducibility are misaligned, according to several speakers (Micah Altman, Lida Anestidou, Andreas Buja, Tim Errington, Irene Qualters, Victoria Stodden, Lawrence Tabak, and Justin Wolfers) and participants. Some of the concerns regard the conflicting messages given to researchers about whether reproducibility research is valued within the community. Researchers are often told that the replications are essential to the health of their scientific community, but most journals do not publish replication papers, most funding agencies do not financially support such work, and many researchers who conduct replication studies can face unpleasant and time-consuming resistance from the community.

Increasing Sharing of Research Material

The increased availability of supplementary research materials, such as data, code, and software, and expanded research methodology descriptions, was highlighted by several speakers (Micah Altman, Ronald Boisvert, Philip Bourne, Tim Errington, Steven Goodman, Randy LeVeque, John Ioannidis, Mark Liberman, Gianluca Setti, Courtney Soderberg, Victoria Stodden, and Justin Wolfers) and participants as being of significant value to enhancing reproducibility. Some journals and funding agencies require data sharing for research publication and funding, but many do not. Furthermore, often the material that is provided in response to these requirements is incomplete, incorrect, or otherwise unreadable.

Enhancing Education and Training

Many speakers (Micah Altman, Chaitan Baru, Yoav Benjamini, Philip Bourne, Xiaoming Huo, and Rafael Irizarry) called for enhanced data science training for people at all levels, including undergraduate and graduate students, beginning

and established researchers, and senior policy leaders. While some of these training courses currently exist and others are being funded by agencies such as the National Institutes of Health, they currently do not sufficiently cover the education landscape; more work needs to be done to identify and fill gaps (Bourne).

ORGANIZATION OF THIS REPORT

Subsequent chapters of this report summarize the workshop presentations and discussion in sequential order. Chapter 2 provides an overview of the importance of reproducibility and discusses two relevant case studies. Chapter 3 focuses on conceptualizing, measuring, and studying reproducibility. Chapter 4 discusses the way forward by using statistics to achieve reproducibility. Finally, Appendix A lists the registered workshop participants, Appendix B shows the workshop agenda, and Appendix C defines acronyms used throughout this report.

2

Overview and Case Studies

The first session of the workshop provided an overview of the importance of reproducibility across scientific communities. Constantine Gatsonis (Brown University, co-chair of the workshop planning committee and chair of the Committee on Applied and Theoretical Statistics) and Giovanni Parmigiani (Dana-Farber Cancer Institute, co-chair of the workshop planning committee) began by introducing the workshop. Lawrence Tabak (National Institutes of Health), Irene Qualters (National Science Foundation), Justin Esarey (Rice University and *The Political Methodologist*), Gianluca Setti (University of Ferrara, Italy, and Institute for Electrical and Electronics Engineers), and Joelle Lomax (Science Exchange) provided perspectives from stakeholders. Victoria Stodden (University of Illinois, Urbana-Champaign) gave an overview of statistical challenges of reproducibility, and Yoav Benjamini (Tel Aviv University) and Justin Wolfers (University of Michigan) discussed reproducibility case studies.

In addition to the references cited throughout this chapter, the workshop planning committee identified the following background references: Alogna et al. (2014); Doyen et al. (2012); Errington et al. (2014); Esarey et al. (2014); Gelman and Loken (2014); Gerber and Green (2000); Harris et al. (2013); Hayes et al. (2006); Hothorn and Leisch (2011); Imai (2005); Johnson et al. (2014); Klein et al. (2014); Molina et al. (2005); Pashler et al. (2012); Rasko and Power (2015); Simons et al. (2014); Stodden et al. (2013a); and Waldron et al. (2014).

OPENING REMARKS FROM THE WORKSHOP CO-CHAIRS

Constantine Gatsonis and Giovanni Parmigiani opened the workshop with a brief overview of the importance of examining reproducibility from a statistical perspective. Gatsonis explained that the main goal of the workshop is to address statistical challenges in assessing and fostering the reproducibility of scientific results by examining three issues: the extent of reproducibility (as measured through quantitative and qualitative approaches), the methodologic causes of reproducibility failures, and the potential methodologic remedies. The methodologic perspective emphasized throughout this workshop is unique and has not been a focus of other current reproducibility discussions, according to Gatsonis. Three overarching questions were examined throughout the workshop:

1. What are appropriate metrics and study designs that can be used to quantify reproducibility of scientific results?
2. How can the choice of statistical methods for study design and analysis affect the reproducibility of a scientific result?
3. Are there analytical and infrastructural approaches that can enhance reproducibility within disciplines and overall?

Gatsonis noted that many researchers believe developing a new conceptual framework for reproducibility is necessary instead of simply cataloging examples in which reproducibility is weak or nonexistent. He hoped that the workshop would start the conversation about such a conceptual framework, as well as how statistical thinking broadly impacts it.

PERSPECTIVES FROM STAKEHOLDERS

Lawrence Tabak, National Institutes of Health

Lawrence Tabak discussed the issue of reproducibility in the biomedical community, specifically from the perspective of the National Institutes of Health (NIH). He stated that reproducibility is a growing challenge, as has been noted by the research community and in multiple publications. The issue crosses research areas but is especially relevant in preclinical research that uses animal models as a prelude to human research, according to Tabak. He noted that science is often viewed as self-correcting and is therefore assumed to be immune from reproducibility problems. In principle, while this remains true in the long term, reproducibility checks in the short and medium term are constrained by interrelated factors.

Insufficient reporting of methodologic approaches is also an issue for a variety of reasons, spanning from limited space within journal articles to fraudulent

claims by researchers, although the latter is the minority. Sena et al. (2007) looked at the prevalence of selected quality characteristics in literature, finding that none of the 600 papers examined provided information on sample-size calculation, very few discussed how randomization was done, and very few blinded the assessment of outcome to the researchers. Tabak emphasized that this does not mean these studies did not look at these issues, but the reader is unsure based on the published record alone.

Another challenge mentioned is the phenomenon of "p-hacking," where data are manipulated (e.g., outliers discarded) until a desired p-value is achieved. The process of p-hacking—which many investigators may not recognize as distorting—can often lead to unsubstantiated correlations being represented as statistically significant. The lack of consideration of sex as a biological variable is also an issue, Tabak explained.

Tabak noted that there are many challenges to ensuring rigor and transparency in reporting science, including incentives to publish positive results and to aim for high-impact factors, poor training, novelty (no negative data), innovation, and grant support. The biomedical research ecosystem should have research integrity at its foundation and balance robust research training (including biostatistics, basic scientific coursework, and experimental design fundamentals) with an environment that rewards networking, mentoring, career development, and collaboration feasibility.

Tabak and NIH Director Francis Collins published a commentary describing NIH's plans to enhance reproducibility, emphasizing that all stakeholders need to be engaged (Collins and Tabak, 2014). NIH has taken a number of steps to raise community awareness, including hosting a June 2014 workshop with journal editors to identify common opportunity areas and a July 2014 workshop with the pharmaceutical trade organization PhRMA to identify common interest with industry; obtaining input from the community on the reagent-related barriers to reproducible research; and participating in meetings with professional societies and institutions. More than 130 journals endorsed the principles discussed at the June 2014 workshop, which were broadly shared in November 2014 through editorials and other notifications.[1] NIH is also engaged in several pilots to address biomedical research and funding issues:

- *Evaluation of scientific premises in grant applications*: new funding opportunities with additional review criteria regarding scientific premises,
- *Checklist and reporting guidelines*: reviewer checklists regarding reporting standards and scientific rigor,

[1] The full list of journals that endorsed these principles is available at National Institutes of Health, 2014, "Endorsements—Principles and Guidelines for Reporting Preclinical Research," http://www.nih.gov/research-training/rigor-reproducibility/principles-guidelines-reporting-preclinical-research.

- *Changes to biosketch*: biosketch pilot with focus on accomplishments instead of just publications,
- *Approaches to reduce "perverse incentives" to publish*: exploration of award options with a longer period of support for investigators,
- *Training*: development of materials discussing the elements of good experimental design, and
- *Other efforts*: use of prize challenges to encourage reproducibility of results and development of a place within PubMed Commons to share and discuss concerns.

Irene Qualters, National Science Foundation

Irene Qualters explained that the National Science Foundation (NSF) has a broad view of reproducibility. She stated that the issues of reproducibility directly impact the credibility and trust afforded to research by both the research community and the public. Research in science and technology has relied on different tools to earn credibility, such as quantification of measurements from experiments, a sustained record of success, and a willingness to retract claims that are demonstrated to be erroneous, limited, or surpassed by new information and data. Thus, progress in science and technology requires that the community acknowledge that any result is to some degree at risk, no matter how carefully it is supported. New evidence, new methods, and new tools always introduce a degree of uncertainty, and science progresses in part by rejecting findings that are disproven over time.

Software is becoming an increasingly critical component within and across all science disciplines. However, software introduces vulnerabilities that are not always appreciated and often challenging to control. Software validation is often employed, especially in areas beyond the research enterprise, such as the complex software environments of nuclear energy and some clinical trials. However, these validation approaches may not be applicable to a foundational research enterprise that often relies on dynamic community software, much of which is contributed by graduate students and postdoctoral fellows. But these communities are building research on the software contributions of others, and their findings need to be credible.

Qualters emphasized the importance of understanding how data are generated and what methodologic approaches were used, as well as what tool employed is crucial for reproducibility. There are powerful tools available to measure reliability and the confidence associated with statistical results, but they are based on assumptions about the underlying data and theory about relationships and causality. As researchers continue to strive to build software to advance science and engineering, an analogous understanding of software tools is needed in order to ensure integrity and identify and measure biases.

Justin Esarey, Rice University and *The Political Methodologist*

Justin Esarey discussed current efforts in political science to improve reproducibility and transparency. He explained that political science is at the forefront in improving research transparency and access to replication resources in the social sciences. For example, many journals require complete replication materials upon acceptance and have done so for years. Some journals even undertake independent verification of results before publication, although the number of journals that do this is small because of resource constraints. The Data Access and Research Transparency initiative has increased the number of journals committed to providing complete, publicly available replication materials for all published work, specifically by

- Requiring authors to ensure that cited data are available at the time of publication through a trusted digital repository (journals may specify which trusted digital repository shall be used);
- Requiring authors to delineate clearly the analytic procedures upon which their published claims rely and, where possible, to provide access to all relevant analytic materials;
- Maintaining a consistent data citation policy that increases the credit that data creators and suppliers receive for their work; and
- Ensuring that journal style guides, codes of ethics, publication manuals, and other forms of guidance are updated and expanded to include strong requirements for data access and research transparency.

The Political Methodologist, the newsletter of the Political Methodology section of the American Political Science Association, recently released an issue[2] focused on reproducibility and transparency problems facing the political science community. One of the tensions identified is that standards for qualitative and quantitative methods can be hard to reconcile (e.g., How can confidentiality for interviewees be ensured? How does one provide replication data for ethnography or for process tracing?). Even defining what replication means and what qualifies as a replication is challenging. Does replication require using the exact same model with the exact same data and the exact same software to generate the same result? Or, does replication check for robustness with slightly different specifications in the same data? Or, does it require verification with independent data, which are often difficult or impossible to get for observational studies? There are also questions about how replication should be integrated into graduate teaching and training, with the

[2] To read this issue, see *The Political Methodologist,* Volume 22, Number 1, Fall 2014, https://thepoliticalmethodologist.files.wordpress.com/2015/02/tpm_v22_n1.pdf.

warning that just requiring the materials necessary for replication materials does not prevent errors. Many other questions remain, in Esarey's view:

- To what extent is proactive error-checking a necessary part of a plan to increase transparency and replicability in the social sciences?
- How should replication projects be rewarded? Should they be published? Should negative replications get more interest?
- What constitutes replication in the case of an analysis of a fixed observational data set for where there exists only one sample (e.g., in time series cross sections (TSCSs) of country data)?
- Where replication is not an option (e.g., in many qualitative methods or observational studies of fixed TSCS data), what would constitute a check on the quality of an empirical model?

Gianluca Setti, University of Ferrara, Italy, and Institute of Electrical and Electronics Engineers

Gianluca Setti explained that the Institute of Electrical and Electronics Engineers (IEEE) is the world's largest professional association dedicated to advancing technological innovation and publishes about 169 journals and magazines. While serving on the IEEE board of directors through 2014, he also chaired a committee on the Future of Information and Convening, which is chartered in part to evaluate the opportunity to promote reproducible research. He emphasized that his presentation was based on his experience and represented his personal opinion.

Setti observed that reproducibility is not necessarily based on just a published paper; it also relies on the availability of data, algorithms, codes, and details of the experimental methods (Buckheit and Donoho, 1995). Because the IEEE deals with many different disciplines, it therefore has to take into consideration many facets of reproducibility, including expectations for documenting proofs, algorithms, and experiments, which may rely on custom-designed equipment (e.g., digital/analog integrated circuit implementation, microelectromechanical systems, nanotechnology, optical devices, etc.). In spite of these challenges, Setti believes that IEEE would have advantages in pursuing improved reproducibility capabilities for collaborations and in making information more visible and directly usable. Improvements would allow researchers to advance technology more easily and practitioners to develop new products faster, and they would reduce the amount of "noise" in the research literature.

Another possible way in which improving reproducibility could benefit IEEE may depend, according to Setti, on making the review process more reliable, making it more difficult to plagiarize a paper, and making it easier to discover false results and avoid retractions. The IEEE Signal Processing Society (SPS) is probably

the most advanced of the IEEE components (Barni and Perez-Gonzalez, 2005; Vandewalle et al., 2009; Piwowar et al., 2007), with all SPS publications explicitly encouraging reproducible research by both instructing researchers to submit all relevant information and providing a digital library that can house supplemental information such as data, algorithms, and code. However, these steps have not yet made a significant impact on reproducibility, according to Setti. He suggested the following steps to improve the situation:

- Create searchable and addressable repositories for data. This would entail at least the following steps:
  - —A unique identifier is needed for each reference data set (perhaps analogous to the Digital Object Identifier system used for other digital information) to facilitate the process of crediting data to authors.
  - —Privacy for sensible data should be guaranteed, which would reduce the liability risk.
  - —The use of standardized sets of data for specific problems should be encouraged.
- Create a cloud-based repository for sharing of codes, algorithms, and circuits, in which each of these elements is stored with its "environment."
  - —Algorithms would run in the cloud using the "original" version of the compiler/program (e.g., C, python, Matlab, Mathematica).
  - —A public-private partnership may be required (e.g., software companies, funding agencies, professional organizations, publishers).
  - —Algorithms should be run, but not revealed, to encourage companies to invest in new reproducibility tools.
  - —Authors could upload algorithms to be linked to the corresponding papers. Metrics need to be developed to track algorithm usage, and algorithms and corresponding papers should be cross-linked.
- Reward positive efforts of authors who contribute to the reproducibility of research.
  - —A well-prepared reproducibility contribution requires time and effort, which currently may be a disincentive due to "publish or perish" pressure.
  - —Publishers could highlight papers that display good practices with regard to replication or which attempt to reproduce earlier work.
  - —Awards could be given for reproducible papers only.
  - —The review of papers with steps that enable replication could be expedited.

Setti noted that there is evidence that papers with reproducibility measures (such as shared data) are more frequently downloaded and cited than papers without these measures. With the rise of citation databases and bibliometrics, there is a strong emphasis on achieving high visibility for publications and this may stimu-

late authors to expend additional resources to ensure the results in their papers are fully reproducible. He suggested that funding agencies might be able to help stimulate reproducibility efforts by requiring reproducibility measures for publicly funded research, which is something that may be possible in the future. Enforcing such a requirement globally, however, may be difficult. It would be beneficial for the reproducibility movement if funding agencies participated in setting up the required infrastructure.

Joelle Lomax, Science Exchange

Joelle Lomax explained that the many systems studied in science, especially in biology, are extraordinarily complex and inherently varied. A key to reproducibility is to understand what variance can be controlled and what variance will always exist in a system. Reproducibility is widely understood to be a problem in both academia (Mobley et al., 2013) and industry (Scott et al., 2008; Perrin, 2014; Steward et al., 2012; Prinz et al., 2011; Begley and Ellis, 2012). However, it is difficult to identify which factors are inhibiting reproducibility and how to go about changing them.

Many hypotheses exist for why experiments fail to replicate, including insufficient reporting of methodology, pressure to withhold negative findings, poor experimental design, biased data manipulation or interpretation, unknown factors that produce variability, and flawed statistical analysis. Lomax said that empirical evidence is needed to determine the basis for this lack of reproducibility.

She explained that Science Exchange and its partners are trying to better understand these challenges as well as how the level of reproducibility (or lack thereof) can be diagnosed. She asserted that the keys to understanding reproducibility are independent replication and transparency. Independent replication is important because biases are known to exist, especially when the person performing the experiment has emotional ties to or economic dependence on the outcome of the result. This bias can lead to flawed data analysis and p-hacking (where data are collected and analyzed in such a way as to achieve a particular p-value). Transparency, specifically openness about exactly how things are being done (e.g., documenting reagents, making all raw and transformed data openly available to other researchers), is also crucial.

Lomax offered a brief summary of Science Exchange, which is an online marketplace and centralized network for more than 1,000 expert providers (both academic core facilities and contract research institutions) offering more than 2,000 experimental services to help perform scientific replications. Science Exchange has utilized this network to undertake independent replication of preclinical research. She explained that Science Exchange has partnered with the Center for Open Science to preregister all experimental designs, protocols, and planned statistical analyses and to share all raw and analyzed data through the Center for Open

Science's Open Science Framework (an integrated project management and data repository).

Preregistration, according to Lomax, allows experts to review key factors such as study design and planned statistical analysis approaches before the data are collected. This is essentially moving the peer-review process to the beginning of the study process to deter the researcher from making certain process adjustments that would inhibit reproducibility.

One of Science Exchange's current projects (partnered with the Center for Open Science and sponsored by the Laura and John Arnold Foundation) is the Reproducibility Project: Cancer Biology, which aims to gather a large data set of replications of preclinical research. The project is replicating the key findings of 50 recent high-impact cancer biology papers though a detailed process, including obtaining input from the original author(s). The findings and methodologies (including protocols and calculations) of these replication studies are prepublished and peer reviewed through *eLife*. Expert independent laboratories that are part of the Science Exchange network are utilized to carry out the replication study, and all protocols, raw data, and results are deposited to the Open Science Framework. Lomax says that this study clarifies what is needed to perform replications of preclinical research and illustrates how difficult it can be to replicate published research.

Lomax concluded by listing some of Science Exchange's other current projects, including partnering with the Prostate Cancer Foundation and PeerJ to look at reproducibility of prostate cancer research, participating in the Reproducibility Initiative, and partnering with reagent companies to validate antibodies.

Panel Discussion

Following their individual presentations, Lawrence Tabak, Irene Qualters, Justin Esarey, Gianluca Setti, and Joelle Lomax participated in a panel discussion with follow-up questions from the audience. The session began with a participant who asked if reproducibility issues were worse today than they were 50 years ago. Qualters noted that the complexity of modeling and simulations has increased considerably over the past 20 years, and this has led to an increase in reproducibility issues. Understanding the uncertainty associated with complex software is a challenge. Tabak noted that science today is much more interdisciplinary than it has been in the past, and this interdisciplinarity makes it more difficult for researchers to truly understand what others are doing. He also noted that the number of publications has increased rapidly, which may also be contributing to the lack of reproducibility.

Another participant noted that while the goal of independent reproducibility is admirable, it assumes that the method being used in the original study is cor-

rect; he wondered what types of approaches exist and what directions should be explored to account for variation. Lomax said that this is being observed in situations in which research plans are preregistered because reviewers of those plans can criticize the original experimental design. She explained that Science Exchange is currently exploring only direct replications to control for as many differences as possible, but additional controls are added if they were not obviously included initially. However, she emphasized that research communities want reproducibility to go even further to encompass the initial variation in results that the participant noted. Esarey agreed that this raises important statistical issues. Many models can be used to analyze a given data set, and it is hoped that a study's results do not vary significantly if one makes minor adjustments to the analysis. Investigators need to understand any such fragility, and statisticians need to develop tools that are robust to minor adjustments, according to Esarey. However, he recognized that using the right methods is inherent in science, and eventually a preferred method is established within a community. Tabak stated that there is a level of practicality that needs to be introduced, particularly in the biomedical community. Replication efforts need to be strategic because resources are limited. NIH is emphasizing the importance of rigorous preclinical research that underlies key decisions before taking that research to human trials.

A participant wondered how free software such as R is influencing current analysis. Esarey said that these types of free open-source software make it less expensive to conduct analyses, and more studies are being carried out because of the abundance of data and analysis tools. However, this free software gives communities more tools to analyze the results from these studies, which is a possible downside because, as mentioned above, some research results can be extracted with certain methods and not others. Setti commented that it is great for research communities that there are free software packages to validate a particular data set, but that often more complex tools are needed. He said that there is an opportunity for open-source developers to partner with software companies to assist the communities.

Another participant asked how funding grants could be used to promote transparency and reproducibility. Tabak said that NIH is piloting new ways of doing grant review. One of these approaches examines the premise of the application, looking at all the work leading up to the application to ensure the research is sound. Currently, NIH only does this indirectly if the preliminary data upon which the application is based come from the principal investigator and his or her team. NIH has also been developing and supporting training modules to stimulate a conversation within the community and to help with graduate and postgraduate training. He said that NIH is considering new funding opportunities for replication studies, as well as options to assess whether preclinical findings should be replicated. Qualters said that embedding reproducibility awareness into a research culture is a big change and it has to take place over time. The many stakeholders

(including individual researchers, research institutions, funding agencies, and the public) need to come together to discuss the best solutions to their community's issues, and this consensus cannot be built overnight. She noted that there are some good resources that discuss incentive issues generally (e.g., Wellcome Trust, 2014).

OVERVIEW OF THE STATISTICAL CHALLENGES OF REPRODUCIBILITY

Victoria Stodden, University of Illinois, Urbana-Champaign

To focus on the framework for the workshop, Victoria Stodden discussed the different ways of viewing and understanding reproducibility. She suggested that it is useful to use one of three modifiers with the word reproducibility: empirical, computational, or statistical. The problems that arise and the remedies for each of these areas are very different. She said that parsing out these differences is essential because people mean different things when they use the term reproducibility.

Empirical reproducibility refers to the traditional notion of scientific reproducibility: the ability to step through the specified physical steps, protocols, and designs as described in a publication. Hines et al. (2014) give an example of empirical reproducibility by describing the difficulty two laboratories faced when collaboratively trying to ensure that both laboratories were producing the same cell signatures in their research from the identical processes. Empirical reproducibility can entail its own special constraints. Stodden shared the example of a 2014 workshop held by the National Academies of Sciences, Engineering, and Medicine's Institute for Laboratory Animal Research that discussed reproducibility issues in research with animals and animal models (NASEM, 2015). That workshop focused mostly on empirical reproducibility, and one of the questions that arose was how to define reproducibility when true replication may mean killing additional animals.

Computational reproducibility, Stodden explained, has only become an issue over the past 20 years. She defines it as any issue arising from having a computer involved somewhere in the work process, from researchers who do bench work and analyze their data with a spreadsheet to researchers doing work on large computing systems with an enormous amount of code and software. Traditionally, going back hundreds of years, Stodden explained, there were two branches of the scientific method: (1) deductive, including mathematics and formal logic, and (2) inductive or empirical, including statistical analysis of controlled experiments. There is now some discussion that expanded computation and the deluge of data are introducing a third (and perhaps a fourth) branch of the scientific method (Donoho et al., 2009): computational, including large-scale simulations and data-driven computational science.

Stodden noted that the scientific community has had hundreds of years to think about the first and second branches of the scientific method but only a couple

of decades to figure out how to use and assess this new technology. Standards for the use of computational modes of inquiry must mature if the technology is to reach its full potential.

This abundance of data and the technology to exploit it have revolutionized a number of scientific fields, according to Stodden. The first change comes from big data and data-driven discovery. High-dimensional data with many more variables than observations are also prevalent, and this poses new challenges of analysis (in contrast to our long history of using analysis to infer meaning from limited data). In Stodden's view, the availability of so much data, and of data-driven discovery, change what it means to carry out inference on the data while still having the ability to reproduce results. The second change that comes from computation is that powerful machines can carry out much more analysis. Elaborate simulations of entire physical systems can be performed, and these calculations can be rerun with a range of parameters so as to explore scientific questions and obtain answers. This was not possible 50 years ago. The third big change noted by Stodden is that deep contributions to science may in some cases only be coded in software, and most of the methodology innovations within the code (e.g., R scripts or more complicated or customized types of software) largely remain inaccessible in the scholarly record.

With respect to the use of computation in conducting research and what it means to be at the standard necessary to establish an acceptable new branch of the scientific method, Stodden believes that the community needs deal with reproducibility issues. She mentioned the following excerpt from Buckheit and Donoho (1995), who paraphrase Jon Claerbout: An article about computational science in a scientific publication is not the scholarship itself, it is merely advertising of the scholarship. The actual scholarship is the complete . . . set of instructions [and data] which generated the figures.

Stodden noted that just reproducing the computational work from a particular study is not the same as replicating the experiments independently, including data collection and software implementation. She stressed that both of these steps are required in order to say that the work has been reproduced. An effort to recode the software and carry out the experiment independently will rarely result in the exact same output; what is most important is being able to tell what was different between the two that led to their distinct outputs. To be able to identify these differences, the researcher who is trying to reproduce the results will need to be able to "step through" the original work and regenerate those results to see what has been done. This is a contribution to science. Both reproducibility and replicability are needed.

A number of infrastructure responses have been created within the scientific community to address the challenges that reproducibility presents, according to Stodden:

- *Dissemination platforms.* These are tools that enable access to codes and data, facilitate statistical verification, and link code data with publications. Examples include ResearchCompendia.org, MLOSS.org, Open Science Framework, IPOL, thedatahub.org, Madagascar, nanoHUB.org, and RunMyCode.org.
- *Workflow tracking and research environments.* These are tools that make it easier for researchers to capture what is important to be communicated, both computationally and physically, from their work. Examples include Vistrails, Galaxy, Pegasus, Kepler, GenePattern, CDE, Sumatra, Jupyter, and Taverna.
- *Embedded publishing.* New modes of publishing go beyond being analogs of papers, with the goal of making it easier to explore and visualize information. Embedded publishing results in more active documents that allow a reader to click on a figure and see it regenerate, see the software, or go somewhere that has the version data set that generated those results. Examples include Verifiable Computational Research, College Authoring Environment, SOLE, SHARE, knitR, and Sweave.

Stodden observed that reproducibility is often a by-product of open-source software, but challenges still exist in documenting this software. Stodden noted that work to strengthen capabilities and practices of computational reproducibility is being done by members of the scientific research community, but it is piecemeal and largely unrewarded as part of their regular jobs.

Statistical reproducibility, which Stodden reminded the audience is the focus of this workshop, covers a range of ways in which statistics and statistical methods influence the degree to which science is reproducible. It includes issues such as the following:

- False discoveries, p-hacking (Simonsohn, 2012), the file drawer problem, overuse and misuse of p-values, and lack of multiple testing adjustments;
- Low power, poor experimental design, and nonrandom sampling;
- Data preparation, treatment of outliers, recombination of data sets, and insufficient reporting and tracking practices;
- Inappropriate tests or models and model misspecification;
- Model robustness to parameter changes and data perturbations; and
- Investigator bias toward previous findings, and conflicts of interest.

Stodden noted that not all of these issues are inherently bad (e.g., having small samples), but they need to be understood as part of the context of an experiment.

She commented that building the cyberinfrastructure to support reproducibility involves bringing together various stakeholders, including researchers,

funding agencies, publishers, scientific societies, and institutions to discuss these issues. In 2009, a roundtable discussion was held at Yale University to discuss pertinent issues for the community, as well as potential remedies and ambitions for the future. The outcome of this workshop was a declaration of reproducible research (Yale Roundtable Participants, 2010). The Institute for Computational and Experimental Research in Mathematics at Brown University held a 2012 workshop[3] to once again bring together stakeholders and continue the discussions. In 2014, the Extreme Science and Engineering Discovery Environment (XSEDE) (an NSF-funded effort that bridges access to high-performance clusters and provides an interface that allows a much broader community to use these resources) held a workshop[4] about what reproducibility means in the high-performance computing context, what the next steps might be, and how we might improve reproducibility in that context.

She summarized three broad reproducibility issues to include the following:

1. Failure of traditional reporting standards to accommodate changes in the research process;
2. Insufficient benchmarking, testing, and validating; and
3. A lack of coordination among research incentives, universities, funding agencies, journals, scientific societies, legal experts and policy makers, internal and ethical pressures, libraries, and the public. She suggested that there are many questions yet to be resolved: What am I doing to make sure that I have a job? What are universities doing? What are promotional and tenure standards? Are people asking about code data, reproducibility, and the robustness of findings? How does funding impact reproducibility? How do legal and policy environments affect the simulations, such as whether or not researchers are generating intellectual property when they develop codes or collect data?

Stodden added that research misconduct obviously impacts reproducibility, although she hopes this is limited in scope within the research community. These situations are complex, but it is important to examine the external pressures facing researchers who are working in a system that rewards outputs without strong regard to their reproducibility.

[3] Institute for Computational and Experimental Research in Mathematics, "Reproducibility in Computational Mathematics: December 10-14, 2012," https://icerm.brown.edu/tw12-5-rcem/, accessed January 12, 2016.

[4] Extreme Science and Engineering Discovery Environment, "An XSEDE Workshop," https://www.xsede.org/web/reproducibility, accessed January 12, 2016.

She concluded by summarizing that the present workshop was designed to focus on statistical reproducibility:

- Are there metrics that can quantify the reproducibility of scientific results?
- How do statistical methods affect the reproducibility of a finding?
- How do routine statistical practices, such as hypothesis test significance thresholds, affect the reproducibility of results?
- Are there analytical and infrastructural approaches that can enhance reproducibility?
- Do we need new frameworks for assessing statistical evidence?
- Do we need the broad adoption of practices for making study protocols, data, code, and workflows openly available?
- How can this be achieved?

She encouraged workshop participants to keep the last question in mind as the discussion of reproducibility continues.

CASE STUDIES

Animal Phenotyping

Yoav Benjamini, Tel Aviv University

Yoav Benjamini began by explaining that while reproducibility and replicability have only recently come to the forefront of many scientific disciplines, they have been prevalent issues in mouse phenotyping research for several decades (Mann, 1994; Lehrer, 2010). He noted that NIH, the National Institute of Neurological Disorders and Strokes, *Nature*, *Science*, and the pharmaceutical industry have recognized that preclinical studies are prone to be nonreplicable. However, he believes there is little agreement about the cause and solution.

Animal phenotyping measures an animal's quantitative and qualitative properties to compare gene strains and mutations and to test drugs and treatments. Animal phenotyping, according to Benjamini, is widespread across many fields of science. One example of animal phenotyping is quantifying exploratory behavior in mice. During drug development, the level of anxiety in mice is documented, often by monitoring how a mouse behaves when it enters a circular arena through a small entrance. A mouse typically explores the arena by moving a small distance into the arena around the perimeter, stopping, and then retreating to the opening. As the mouse becomes more familiar with the arena, it goes a small distance farther around the perimeter than on the first attempt, stops, and again retreats to the entrance. This behavior pattern continues until the mouse eventually makes it all the

way around the perimeter and back to the entrance. This behavior can be tracked automatically, analyzed statistically, and quantified in a number of ways (e.g., the total number of steps the mouse takes, the maximum speed, the percent of time spent in the center of the arena, etc.). The percent of time spent in the center of the arena is a popular measure of anxiety; however, there is a lack of standardization regarding the time, the duration, the arena's size and shape, and the mouse breed. This lack of standardization may cause serious replicability problems (*Nature Methods*, 2012; Funio et al., 2012).

The lack of standardization is a serious problem that scientific communities should care about, Benjamini emphasized. He discussed a Crabbe et al. (1999) study where an experiment was conducted in three laboratories with strict standardization variables. This study found there were genotype and laboratory effects that may yield results that are idiosyncratic to a particular laboratory. Additional studies showed that differences between laboratories could contribute to failures to replicate results of genetic experiments (Wahlsten, 2001).

Later experiments (Richter et al., 2009, 2011; Würbel et al., 2013) compared two strains of mice (C57 and CBA) in six commercial and academic laboratories across Europe and showed random laboratory and interaction effects. He explained that the particular laboratory effect in a new laboratory is unknown but its random effect can be eliminated in this case by design by running the two strains in the same laboratory. However, the particular genotype *x* laboratory (GxL) mixed-model interaction cannot be eliminated by design.

Initially, the existence of significant GxL interaction was considered a lack of replicability, but Benjamini argued that the existence of GxL interaction cannot be avoided in part because the genotyping by laboratory effect is unknown. He explained that some aspects can be automated and made more uniform; however, some researchers have raised concerns about whether increased uniformity in laboratories will heighten the effects of small unimportant differences between laboratories, which could harm replicability rather than improve it.

Benjamini stated that if the GxL variability $\left(\sigma^2_{GxL}\right)$ were known and could be estimated, then it could easily be corrected for by subtracting the mean of one strain of mice from the mean of another strain and dividing this by the regular t test $\left(\sigma^2_{Within}(1/n+1/n)\right)$ plus the variability $\left(2\sigma^2_{GxL}\right)$, making the calculation

$$\frac{Mean\left(M_{G1}\right)-Mean\left(M_{G2}\right)}{\left(\sigma^2_{Within}\left(\frac{1}{n}+\frac{1}{n}\right)+2\sigma^2_{GxL}\right)^{1/2}}$$

Benjamini argued that the interaction size is the correct measure to compare genetic difference when examining replicability across laboratories. He emphasized that design quality, large sample size, and transparency alone will not solve the

issue of replicability. Increasing the sample size will diminish $\left(\sigma^2_{Within}(1/n+1/n)\right)$, but it will not help with the variability. He stated that many nonreplicable results could be screened out if laboratories knew the GxL variability. However, he said the proportion of errors committed by falsely rejecting null hypotheses (known as the false discovery rate) is still high but could be approached using a false discovery rate controlling method (Benjamini and Hochberg, 1995). Therefore Benjamini recommends reporting both a p-value and a GxL-adjusted p-value for replicability, or perhaps a ratio of the two.

The GxL variability can be estimated by using batch variability (e.g., litters of animals, days the experiment was conducted) as a surrogate for GxL variability or by purposely injecting variation into the experiment's environmental condition when conducted in a single laboratory (Richter et al., 2009). The latter heterogenizing approach is controversial because it is the opposite of standardizing, but Benjamini believes this approach allows researchers to control the known variables. To improve estimates of GxL variability, Benjamini proposed making use of large publicly available databases of mouse phenotyping results (e.g., the International Mouse Phenotyping Consortium[5]).

He concluded by reiterating Victoria Stodden's point that reproducibility of computing problems is difficult. He also noted that there is a replicability problem with two key statistical sources in animal phenotyping: using an inappropriate measure for variability (e.g., ignoring GxL interaction) and failing to adjust for selective inference. He stated that transparency, standardization, and large sample sizes do not eliminate these problems in single-laboratory experiments, but community efforts can help. He believes that replicability can move from a burden to an asset by estimating the GxL variability from multilaboratory experiments to evaluate newly suggested measurement tools and devices.

A participant asked about the infrastructure developed to address GxL interactions. Benjamini explained that the infrastructure is not well developed yet, but he and others are trying to build a database with the cooperation of the International Mouse Phenotyping Consortium, with the hopes of getting scientists to contribute results to help enrich the system.

Another participant asked if interlaboratory variability should be treated as variation in bias or whether there is something interesting about anxiety, perhaps the genetic source of anxiety in the interlaboratory variability. Moreover, the participant wondered if there is something to be learned about anxiety from knowing the ways in which laboratories are different. Benjamini said that there have been many efforts to try to identify and standardize sources of variability. However, there are often far too many sources of variability to standardize everything effectively,

[5] The International Mouse Phenotyping Consortium website is http://www.mousephenotype.org, accessed January 8, 2016.

so effort is taken to standardize the variables that are believed to be important and report the other variables. This approach gives researchers an opportunity to investigate different variables in follow-up studies and demonstrates why transparency is an important issue.

A participant questioned the generalizability of the approach, specifically if the approaches to analyzing, modeling, and judging reproducibility of mice data apply to other research areas or if a different framework would need to be developed. Benjamini responded that this could be generalized by identifying the real uncertainty (Mosteller and Tukey, 1977) because the uncertainty between the animals is less relevant than the interaction between genotypes and laboratories. He stressed that insisting on the correct relative level of the uncertainty is an important issue that can be generalized to other areas and other fields.

A participant wondered if the laboratories would get the same results if they exchanged animals. Even if the mice are genetically identical, Benjamini responded, they may have diverged because of evolution in the cage, and they could be epigenetically different because they may have different methylation statuses of the DNA from the different environments. The participant noted it would be interesting to know if results were consistent when moving the mice or whether the laboratory of origin or the environment had changed it. Benjamini said that animal exchanges happen and significant variation has been shown among supposedly genetically identical mice (Kiselycznyk and Holmes, 2011), but a homogeneous set of animals is typically used for drug discovery and experimentation. He suspects this is not a large problem within the research.

Another participant commented that attempting to measure many variables about the laboratories and enter them into the model first, assuming the treatment is the same and neither bias nor fraudulent behavior is an issue, may be more efficient for recovering results and measuring the source of the variation. Benjamini responded that when assessing replicability, the concern is not what happened in the current experiment but whether another laboratory would get the same results using the described methodology. He cautioned against overspecifying properties of the laboratory because the results may end up being more particular to the laboratory. The goal is to estimate the interaction because it is the best assurance of what will happen when the next laboratory tries to replicate these results. Benjamini suggested either using a mixture of the fixed- and mixed-model analyses to capitalize on the strengths of both or simply reporting both. This means doing the best analysis possible from the two laboratories and then estimating the adjusted variability so future researchers know what to expect.

Capital Punishment

Justin Wolfers, University of Michigan

Justin Wolfers introduced his case study, which illustrates some of the challenges of doing empirical research and trying to reassess existing literature. Wolfers's discussion was based on work he did with John Donohue, a law professor at Stanford, which examined the deterrent effect of the death penalty (Donohue and Wolfers, 2006). Wolfers noted that because the death penalty is a politically contentious issue, the challenges associated with performing objective social science research are amplified.

The social, legal, economic, and political histories of the death penalty in the United States are integrally wrapped up in the evolution of social science, according to Wolfers. Figure 2.1 shows the history of executions in the United States. In the early 1900s, approximately 150 people were executed each year in the United States. From 1907 to 1917, nine states abolished capital punishment, which decreased the total number of executions for the country during that time. During the 1950s and 1960s, there was further decline in public support of the death penalty. Sellin (1959) looked at what happened in a state that had the death penalty versus in an adjoining state that did not; his study suggested there was no deterrent effect of the death penalty. The laws during this time did not change, but the death penalty was rarely used; the total number of executions plummeted to zero between 1967 and 1971, despite the fact that many states still had it on the books and there were still people imprisoned for capital murder.

In 1972, the Supreme Court deemed all existing death penalty statutes in the United States unconstitutional (Furman v. Georgia), which led to a de facto moratorium on capital punishment. In contrast to Sellin's study, Ehrlich published a paper in 1975 that analyzed the connection between executions and homicides and concluded that for each person executed, eight future homicides were prevented. The Supreme Court cited Ehrlich's paper in its 1976 decision (Gregg v. Georgia) that allowed capital punishment to resume. In 1978, the National Research Council released a report that assessed the research cited by the Supreme Court and found that there was little evidence from social science to suggest that executions deterred homicide (NRC, 1978). In spite of the Supreme Court ruling, the death penalty fell largely into disuse with just a few executions per year.

Then in the 1980s, a few large states began to express interest in using the death penalty again, and the economic and political debates were reignited. Wolfers said that various scholars, many of whom were located in the southern United States, started to write research papers looking at the relationship between execution and homicide rates and once again claimed that higher execution rates led to lower homicide rates.

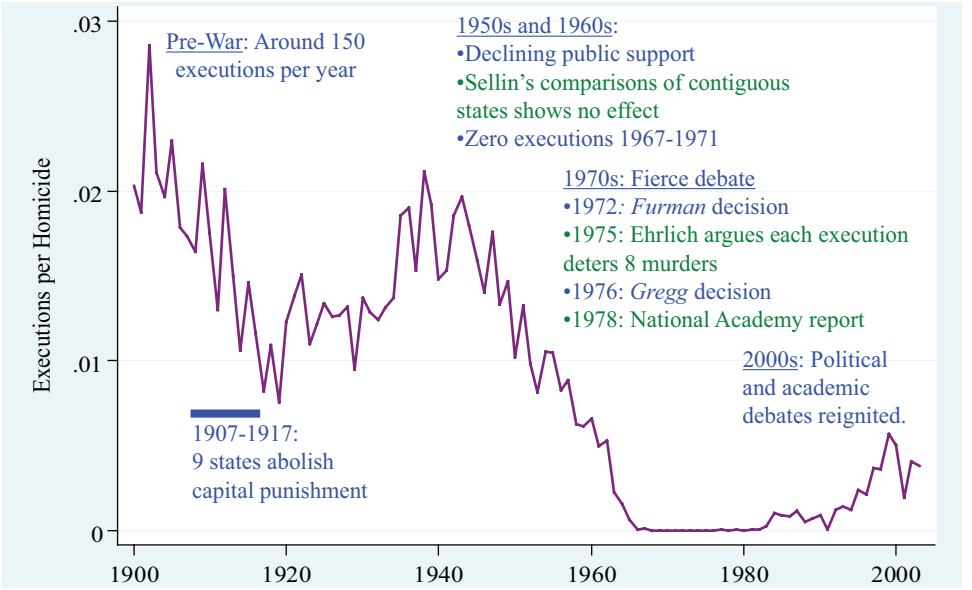

FIGURE 2.1 A century of executions in the United States on an executions-per-homicides basis. SOURCE: Courtesy of Justin Wolfers, University of Michigan, presentation to the workshop.

In the late 1990s and early 2000s, most of the economics literature suggested that executions deterred homicide (e.g., Dezhbakhsh and Shepherd, 2006; Dezhbakhsh et al., 2003; Shepherd, 2005; Mocan and Gittings, 2003; Zimmerman, 2004; Cloninger and Marchesini, 2001). However, Wolfers pointed out that one study, conducted by researchers who focused on prison conditions rather than the death penalty (Katz et al., 2003), controlled for what was happening with the death penalty and found that there was no deterrent factor. Indeed, they inferred that there was a slight incentive to commit murder. The deterrent effect shown in literature from this time period is outlined in Table 2.1.

Wolfers noted that the broader context is one of robust debate in the policy world. Texas decided during this time to ramp up use of executions, while California sentenced many people to death row but executed few of them. Several state governors, including those in Illinois and New Jersey, instituted moratoriums on the death penalty.

He described some of the strong claims being made, such as Emory University law professor Joanna Shepherd's statement before Congress that there is a "strong consensus among economists that capital punishment deters crime," "the studies are unanimous," and "there may be people on the other side that rely on older

TABLE 2.1 The State of the Capital Punishment and Deterrence Literature, circa 2008

Literature	Data Used	Number of Lives Saved per Execution
Dezhbakhsh and Shepherd (2006)	1960-2000 state panel	8
Dezhbakhsh et al. (2003)	1977-1996 county panel	18
Shepherd (2005)	1977-1999 monthly state panel	2
Mocan and Gittings (2003)	1984-1997 state panel	4
Zimmerman (2004)	1978-1997 state panel	14
Cloninger and Marchesini (2001)	Illinois 2000-2003 moratorium Texas April 1996-April 1997 moratorium	40 18
Katz et al. (2003)	1950-1990 state panel	–0.8

SOURCE: Courtesy of Justin Wolfers, University of Michigan, presentation to the workshop.

papers and studies that use outdated statistical techniques or older data, but all of the modern economic studies in the past decade have found a deterrent effect. So I'm not sure what other people are relying on" (Terrorist Penalties Enhancement Act of 2003). Wolfers added that this issue came up in political debates as well. In the 2000 presidential election, George W. Bush supported the use of the death penalty because it "saves other people's lives" (Commission on Presidential Debates, 2010), while Barack Obama stated that the death penalty "does little to deter crime" (Obama, 2006).

Wolfers noted that research intending to analyze the deterrent effect of the death penalty could be approached by measuring a causal effect of an experiment with subjects. The standard error of this analysis would be large or small depending on whether the number of subjects considered was small or large, respectively. If all experiments are being reported and illustrated using a funnel chart (see Figure 2.2a), then a relationship is illustrated where the larger the standard error, the more variable the estimates across different experiments. The smaller the standard error, the less variable the findings should be, with 5 percent falling outside the 95 percent band of the funnel. If instead scholars choose to report only the results that showed statistically significant evidence of a deterrent effect of the death penalty, the literature would be limited to the analyses that fall above the 95 percent band of the funnel (see Figure 2.2b). Wolfers explained that this leaves a footprint in the data that shows a correlation between how large the estimates of the effect of the death penalty are and the standard error of the particular studies.

He took the key measurements from some of the studies in the death penalty literature and graphed their results, where the horizontal axis is the standard error and the vertical axis is the effect of each execution on the number of homicides.

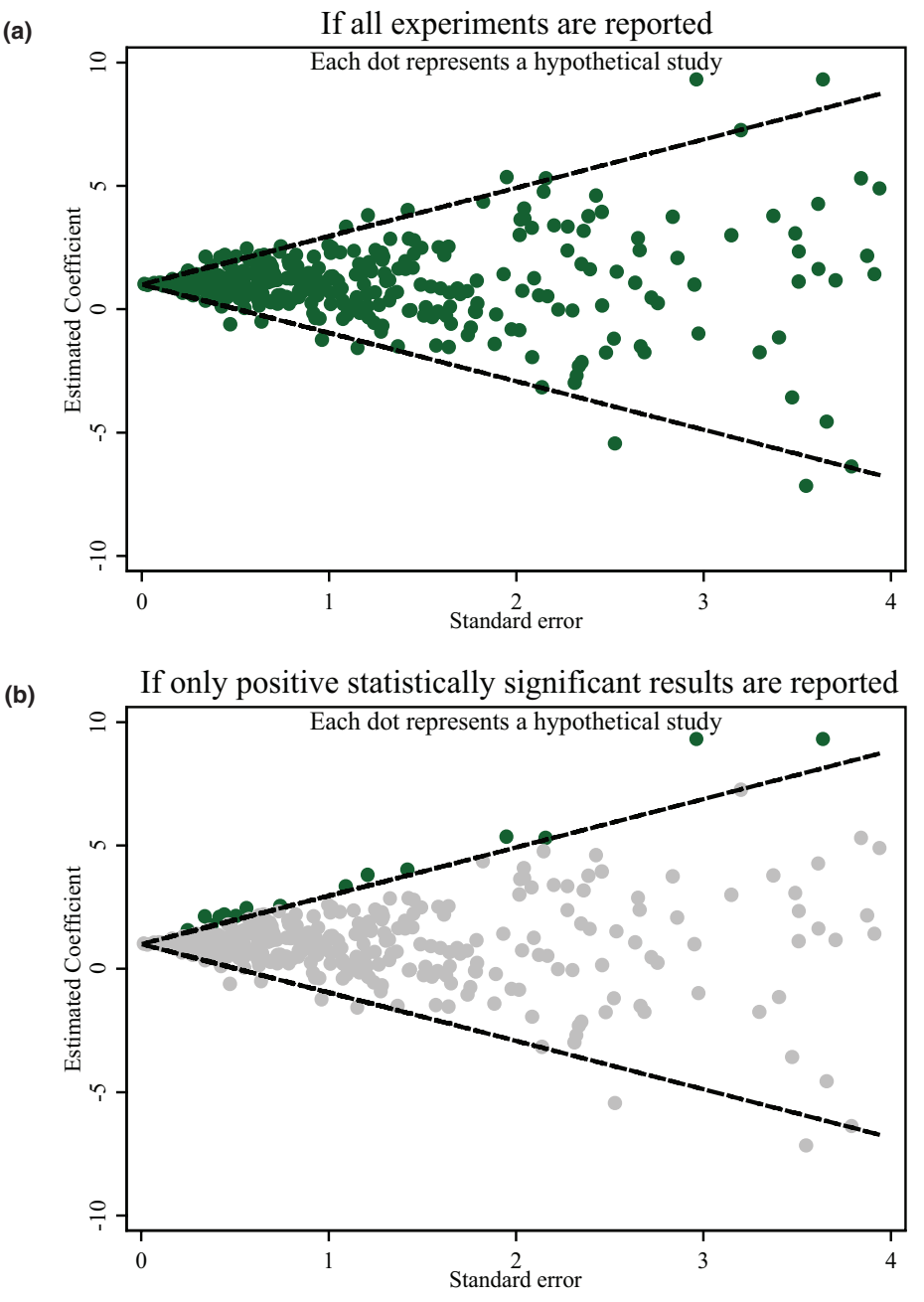

FIGURE 2.2 Funnel charts showing estimated coefficients and standard error if (a) all hypothetical study experiments are being reported and (b) if only statistically significant results are reported. SOURCE: Courtesy of Justin Wolfers, University of Michigan, presentation to the workshop.

Figure 2.3 shows the central estimate from each study, which Wolfers explained showed a clear, statistically significant relationship with the six observations. He noted that imprecise results are reported when they yield extremely large effects.

In Wolfers's view, the norm in most social science work is to subject the main estimates to a variety of sensitivity testing—for example, to see what happens if a slightly different regression specification is used, the functional form is altered, a couple of states are left out, or the sample period is changed. He explained that the literature shows a robust relationship between the size of the estimate and the size of the standard error, with the exception of the Katz et al. (2003) paper. Wolfers argued that this implies selective reporting within the literature. This could be a type of specification search or file drawer problem where the central estimates are not representative of the underlying distribution of estimates that one could find from a cleaner look at the data.

Wolfers explained that when he and Donohue replicated each of these studies, they found coding errors and what they believed were fairly substantial judgments that could not be supported. They concluded that none of the studies that showed large deterrent effects were convincing. He argued that replication research and

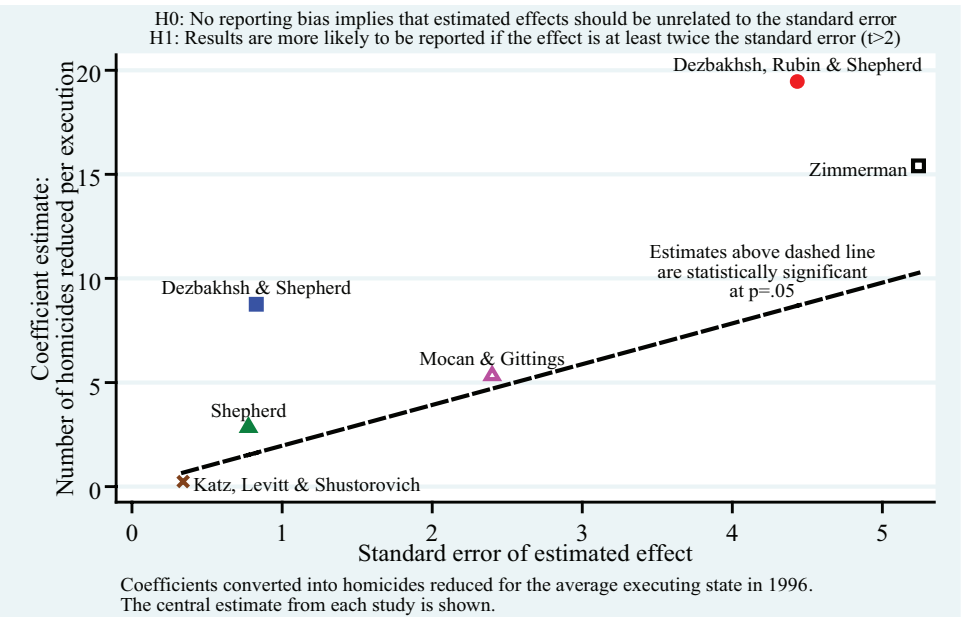

FIGURE 2.3 Reporting bias of estimated effects of executions on homicide. NOTE: H0, No reporting bias implies that estimated effects should be unrelated to the standard error; H1, Results are more likely to be reported if the effect is at least twice the standard error. SOURCE: Courtesy of Justin Wolfers, University of Michigan, presentation to the workshop.

sensitivity testing like this work is important because it begins to fill in missing pieces and identify unconvincing evidence, essentially finding the information that had not been recorded.

Responding to Isaac Ehrlich's study (1975) that found each execution deterred eight homicides, Wolfers commented that it is hard to find a convincing and reliable correlation using the data and statistical technique from that study. The study's data (from 1933 to 1967) showed that both homicides and executions decreased, and Wolfers pointed out that time series analysis would thus imply that decreasing the number of executions reduces homicides. However, Ehrlich had specified the model in terms of the log number of executions and, as the number of executions falls, the log number of executions shows a positive correlation with the decrease in homicides. However, this analysis does not accurately capture the increases in homicides that occurred during some years because using the log smooths the results; if data from 1960 to the early 2000s are analyzed instead (as is done by Dezhbakhsh and Shepherd, 2006), the small decline of executions in 1960 coincides with a huge rise in homicides that lasts throughout the 1980s. Wolfers noted that this also coincides with the crack cocaine epidemic. Executions stayed low while homicides rose for the next 20 years. As executions began to rise slightly (while remaining low overall) around 1990, the data show a massive deterrent effect on the homicide rate. However, Wolfers questioned why tiny changes in executions have massive effects on homicides but massive changes in executions do not have massive effects on homicides. Because the time frame chosen for the analysis resulted in vastly different results, he concluded that simple statistical misuse (such as taking logs around zero) could dramatically alter findings.

Wolfers then discussed the natural experiment period between 1972 and 1976 when the United States no longer had the death penalty. The homicide rate rose throughout that time, implying that eliminating the death penalty was to blame. However, homicide rate data from Canada closely mirrors the pattern seen in the United States during that period (although the Canadian homicide rate is only one-third that of the United States). Similarly, Canada eliminated the death penalty in 1965, and the homicide rate there rose. However, the rate rose similarly in the United States. He also noted that when the Supreme Court eliminated the death penalty in 1973, there was obviously no change in the number of executions in the 10 states that had already abolished the death penalty. However, homicide rates also rose in those states. Wolfers stressed the importance of comparison groups when doing analyses such as this from observational data.

Dezbakhsh and Shepherd (2006) analyzed state homicide rate data 3 years before and after the death penalty was abolished in each state. They found that abolishing the death penalty caused the homicide rate to rise considerably, while its reinstatement led to a decline in the rate. However, Wolfers noted that this analysis

misses relevant comparisons during these time periods, as well as the fact that rates were similar across other states during the years being examined.

In Wolfers's view, many of these papers on the death penalty also suffer from previously mentioned serious limitations, such as coding errors that eliminate the wrong observations (Mocan and Gittings, 2003) and judgment calls that skew the findings from the analysis (Dezhbakhsh et al., 2003).

Wolfers cautioned that unpleasant confrontations are a major disincentive to attempting to replicate others' work. For example, when Donohue and Wolfers (2005) were published, several authors of the work that was replicated sharply criticized the new paper, though they did not provide the original data to allow a more thorough examination of their analysis. Donohue and Wolfers (2005) ultimately concluded that there are insufficient data to determine the influence of the death penalty while noting that the existing literature reflects problems including publication bias; neglect of comparison groups; coding errors; highly selected samples, functional forms, regressors, and samples; and overstatements about statistical significance.

In 2012, the National Research Council released a report on the evidence of the deterrent influence of the death penalty. That report stated, "These studies should not be used to inform judgments about the effect of the death penalty on homicide, and should not serve as a basis for policy decisions about capital punishment. . . . Fundamental flaws in the research we reviewed make it of no use in answering the question of whether the death penalty affects homicides rates" (NRC, 2012). The broader lessons of this case study, according to Wolfers, are as follows: (1) the politically charged nature of this debate may change the discussion and influence the analysis; (2) selective reporting is a pervasive problem in many attempts to look at funnel plots with respect to social science research; (3) there is a great need for sensitivity testing for replication; (4) improved norms and definitions are needed in social science research; (5) there is an important role for impartial referees; and (6) there are substantial disincentives for researchers getting involved in this research.

In response to a participant's question, Wolfers suggested that the community work together to lay out a set of definitions about replication so it is clear what was done when someone claims to have replicated research. At present, "replication" can mean anything from simply making the code work on a different computer to thoroughly evaluating and validating the underlying analysis. He suspects one of the difficulties will be coming up with a set of definitions across disciplines.

Another participant stated that social science literature, especially when it gets into the newspapers, often has confounders and flawed results that can be attributed to poor data analysis or fitting models without thinking about the available data. This participant wondered if there are policies that could be implemented to improve the data analysis. Wolfers responded that although review is imper-

fect, a better system is not currently available. The flaws he identified could have been spotted during peer review, but mistakes happen even in reputable journals. For example, because codes are often not shared, coding errors cannot be found. Wolfers suggested that the scientific process could be improved by requiring that every published paper include the archived data, but this does not address the possibility that an archived code might fail to run or deliver the expected results. Some journals employ a research assistant to run the code to make sure that the numbers came out one time on one machine, but this is expensive. Wolfers also commented that in some top-tier newspapers, the refereeing criteria are sometimes more rigorous than those used in peer review.

Often, scientific reporting is skewed by university press offices that draft press releases in more absolute and less nuanced ways than researchers would. Andrew Gelman (2015) has suggested that researchers should be held accountable for exaggerations in press releases about their work; Wolfers agrees that this would improve the accuracy of these releases. Another participant commented that the misportrayal of science in the media is not unique to the social sciences. This participant sent a note to a top-tier journal pointing out that the results of a paper being used to guide ovarian cancer treatment were based on confounding experimental design. Stating that this letter was too technical for their readership, the journal refused to publish it; however, *The New York Times* published the letter two weeks later.

Wolfers noted that in economics, and in the social sciences more broadly, there is very little demand for replication work. Donohue and Wolfers (2005) overcame this by examining several major papers in the existing literature and systematically proving the results to be false. He believes that if they instead had examined only one paper it would not have been publishable. He suggested that there is actually a disincentive for journals to publish replicability research because it requires time and energy to be as careful as they need to be; good replication is incredibly valuable, but bad replication is destructive.

In response to a question by a participant regarding the benefit of occasionally calling out severe mistakes as fraud, Wolfers claimed he tends to avoid using this word because even researchers doing indefensible work can change their ways and adopt better practices in the future. In response to another participant, he clarified that having flawed analysis and making up data require a different level of response, and often the former does not rise to the level of fraud.

Another participant asked what could be done when the data cannot be shared due to data use agreement, as is often the case with social science settings (e.g., health policy). Wolfers noted that journals often confront this issue, and some have instituted a policy requiring researchers to submit all the code needed to replicate the analysis with the restricted data. Since access to the data is rarely granted, the editors have to trust that this code would actually replicate the work. This approach makes it clear how the researcher approached the analysis.

A participant posed a final question about what the scientific community could do to communicate its concerns and possible solutions to policy makers and stakeholders. Wolfers offered that in the case of the death penalty analysis, he was relieved that the Academies used its role as a highly respected independent referee while examining the literature. However, he noted that this might not be the best use of resources in all cases. He said that the social sciences are moving toward being more open through blogs that discuss the research process, which foster more public and transparent discussion of reproducibility issues. The success of this approach depends on the culture of each discipline. Institutionally, Wolfers noted that many funders are excited about developing a replicability standard, and much of the movement in the natural sciences is coming from the funders insisting that data be in the public domain.

3

Conceptualizing, Measuring, and Studying Reproducibility

The second session of the workshop provided an overview of how to conceptualize, measure, and study reproducibility. Steven Goodman (Stanford University School of Medicine) and Yoav Benjamini (Tel Aviv University) discussed definitions and measures of reproducibility. Dennis Boos (North Carolina State University), Andrea Buja (Wharton School of the University of Pennsylvania), and Valen Johnson (Texas A&M University) discussed reproducibility and statistical significance. Marc Suchard (University of California, Los Angeles), Courtney Soderberg (Center for Open Science), and John Ioannidis (Stanford University) discussed assessment of factors affecting reproducibility. Mark Liberman (University of Pennsylvania) and Micah Altman (Massachusetts Institute of Technology) discussed reproducibility from the informatics perspective.

In addition to the references cited in this chapter, the planning committee would like to highlight the following background references: Altman et al. (2004); Begley (2013); Bernau et al. (2014); Berry (2012); Clayton and Collins (2014); Colquhoun (2014); Cossins (2014); Donoho (2010); Goodman et al. (1998); Jager and Leek (2014); Li et al. (2011); Liberati et al. (2009); Nosek et al. (2012); Peers et al. (2012); Rekdal (2014); Schooler (2014); Spence (2014); and Stodden (2013).

DEFINITIONS AND MEASURES OF REPRODUCIBILITY

Steven Goodman, Stanford University School of Medicine

Steven Goodman began his presentation by explaining that defining and measuring reproducibility is difficult; thus, his goal was not to list every measure but rather to provide an ontology of ways to think about reproducibility and to identify some challenges going forward. When trying to define reproducibility, Goodman noted, one needs to distinguish among reproducibility, replicability, repeatability, reliability, robustness, and generalizability because these terms are often used in widely different ways. Similarly, related issues such as open science, transparency, and truth are often ill defined. He noted that although the word "reproducibility" is often used in place of "truth," this association is often inaccurate.

Goodman mentioned a 2015 Academy Award acceptance speech in which Julianne Moore said, "I read an article that said that winning an Oscar could lead to living five years longer. If that's true, I'd really like to thank the Academy because my husband is younger than me." According to Goodman, the article that Moore referenced (Redelmeier and Singh, 2001) had analysis and data access issues. He noted that an examination of some of these issues was later added to the discussion section of the paper and when the study was repeated (Sylvestre et al., 2006), no statistically significant added life expectancy was found for Oscar winners. In spite of the flawed original analysis, it is repeatedly referenced as fact in news stories. Goodman said this is a testament to how difficult it is to rectify the impact of nonreproducible findings.

Many philosophers and researchers have thought about the essence of truth over the past 100 years. William Whewell, Goodman noted, was one of the most important philosophers of science of the 19th century. He coined many common words such as scientist, physicist, ion, anode, cathode, and dielectric. He was an influential thinker who inspired Darwin, Faraday, Babbage, and John Stuart Mill. Goodman explained that Whewell also came up with the notion of consilience: that when a phenomenon is observed through multiple independent means, its validity may be greater than could be ascribed to any one of the individual inferences. This notion is used in the Bradford Hill criteria for causation[1] (Hill, 1965) and in a range of scientific teaching. Edwin Wilson reenergized the notion of consilience in the late 1990s (Wilson, 1998). Goodman explained that the consilience discussions are in many ways a precursor to the current discussions of what kind of reproducibility can be designed and what actually gets science closer to the truth.

[1] A group of minimal conditions is necessary to provide adequate evidence of a causal relationship between an incidence and a possible consequence.

Ronald Aylmer Fisher, Goodman explained, is one of the most important statisticians of the 20th century. Fisher wrote, "Personally, the writer prefers to set a low standard of significance at the 5 percent point. . . . A scientific fact should be regarded as experimentally established only if a properly designed experiment rarely fails to give this level of significance" (1926). Goodman stated that if this were how science was practiced by even by a small percentage of researchers, there would be significant progress in the realm of reproducibility.

Goodman noted that it is challenging to discuss reproducibility across the disciplines because there are both definitional and procedural differences among the various branches of science. He suggested that disciplines may cluster somewhat into groups with similar cultures as follows:

- Clinical and population-based sciences (e.g., epidemiology, clinical research, social science),
- Laboratory science (e.g., chemistry, physics, biology),
- Natural world-based science (e.g., astronomy, ecology),
- Computational sciences (e.g., statistics, applied mathematics, computer science, informatics), and
- Psychology.

Statistics, however, works in all sciences, and most statisticians cross multiple domain boundaries. This makes them ideal ambassadors of methods and carriers of ideas across disciplines, giving them the scope to see commonalities where other people might not.

Goodman noted that while some mergers are occurring—such as among genomics, proteomics, and economics—a few key differences exist between these communities of disciplines:

- *Signal-measurement error ratio.* When dealing with human beings, it is difficult to increase or decrease the error simply by increasing the sample size. In physics, however, engineering a device better can often reduce error. The calibration is different across fields and it results in a big difference in terms of what replication or reproducibility means.
- *Statistical methods and standards for claims.*
- *Complexity of designs/measurement tools.*
- *Closeness of fit between a hypothesis and experimental design/data.* In psychology, if a researcher is studying the influence of patriotic symbols on election choices, there are many different kinds of experiments that would fit that broad hypothesis. In contrast, a biomedical researcher studying the acceptable dose of aspirin to prevent heart attacks is faced with a much narrower range of experimental design. The closeness of fit has tremendous

bearing on what is considered a meaningful replication. While in many disciplines a language of direct replication, close replication, and conceptual replication exists, that language does not exist in biomedicine.

- *Culture of replication, transparency, and cumulative knowledge.* In clinical research and trials, it is routine to do studies and add their results. If one study is significant but the next one is not, neither is disconfirmed; rather, all the findings are added and analyzed together. However, in other fields it would be seen as a crisis if one study shows no effect while another study shows a significant effect.
- *Purpose to which findings will be put (consequences of false claims).* This is something that lurks in the background, especially with high-stakes research where lives are at risk. The bar has to be very high, as it is in clinical trials (Lash, 2015).

Goodman highlighted some of the literature that describes how reproducible research and replications are defined. Peng et al. (2006) defined criteria for reproducible epidemiology research (see Table 3.1), which state that replication of results requires that the analytical data set, the methods of the computer code, and all of the necessary metadata in the documentation necessary to run that code be available and that standard methods for distribution be used. Goodman explained that, in this case, *reproducible research* was research where you could see the data, run the code, and go from there to see if you could replicate the results or make adjustments. These criteria were applied to computational science a few years later (Peng, 2011), as shown in Figure 3.1. In this figure, the reproducibility spectrum progresses from code only, to code and data, to link and executable code and data. Goodman noted that this spectrum stops short of "full replication," and Peng stated that "the fact that an analysis is reproducible does not guarantee the quality, correctness, or validity of the published results" (2011).

Goodman gave several examples to illustrate difficulties in reproducibility, including missing or misrepresented data in an analytical data set and limited access to important information not typically published or shared (such as case study reports in the case of Doshi et al. [2012]).

The National Science Foundation's Subcommittee on Robust Research defined the following terms in their 2015 report on reproducibility, replicability, and generalization in the social, behavioral, and economic sciences:

- *Reproducibility.* "The ability of a researcher to duplicate the results of a prior study using the same materials . . . as were used by the original investigator. . . . A second researcher might use the same raw data to build the same analysis files and implement the same statistical analysis . . . [in an attempt to] yield the same results. . . . If the same results were not obtained,

TABLE 3.1 Criteria for Reproducible Epidemiologic Research

Research Component	Requirement
Data	Analytical data set is available.
Methods	Computer code underlying figures, tables, and other principal results is made available in a human-readable form. In addition, the software environment necessary to execute that code is available.
Documentation	Adequate documentation of the computer code, software environment, and analytical data set is available to enable others to repeat the analyses and to conduct other similar ones.
Distribution	Standard methods of distribution are used for others to access the software, data, and documentation.

SOURCE: Peng et al. (2006).

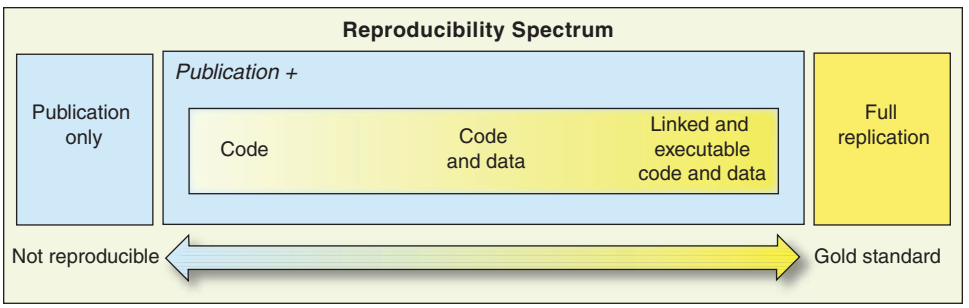

FIGURE 3.1 Spectrum of reproducibility in computational research. SOURCE: Republished with permission of *Science* magazine, from Roger D. Peng, Reproducible research in computational science, *Science* 334(6060), 2011; permission conveyed through Copyright Clearance Center, Inc.

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" (NSF, 2015).

- *Replicability.* "The ability of a researcher to duplicate the results of a prior study if the same procedures are followed but new data are collected . . . a failure to replicate a scientific finding is commonly thought to occur when one study documents [statistically significant] relations between two or more variables and a subsequent attempt to implement the same operations fails to yield the same [statistically significant] relations" (NSF, 2015). Failure to replicate can occur when methods from either study are flawed

or sufficiently different or when results are in fact statistically compatible in spite of differences in significance. Goodman questioned if this is an appropriate and sufficient definition.

Goodman referenced several recent and current reproducibility efforts, including the Many Labs Replication Project,[2] aimed at replicating important findings in social psychology. He noted that replication in this context means repeating the experiment to see if the findings are the same. He also discussed the work of researchers from Bayer who are trying to replicate important findings in basic science for oncology; Prinz et al. stated the following:

> To substantiate our incidental observations that published reports are frequently not reproducible with quantitative data, we performed an analysis of our early (target identification and validation) in-house projects in our strategic research fields of oncology, women's health and cardiovascular diseases that were performed over the past 4 years. . . . In almost two-thirds of the projects, there were inconsistencies between published data and in-house data that either considerably prolonged the duration of the target validation process or, in most cases resulted in termination of the projects because the evidence that was generated for the therapeutic hypothesis was insufficient to justify further investment in these projects. (2011)

Goodman pointed out that "reproducible" in this context denotes repeating the experiment again. C. Glenn Begley and Lee Ellis found that only 6 of 53 landmark experiments in preclinical research were reproducible: "When findings could not be reproduced, an attempt was made to contact the original authors, discuss the discrepant findings, exchange reagents and repeat experiments under the author's direction, occasionally even in the laboratory of the original investigator" (2012). Sharon Begley wrote about C. Glenn Begley's experiences in a later *Reuters* article: [C. Glenn] Begley met for breakfast at a cancer conference with the lead scientist of one of the problematic studies. "We went through the paper line by line, figure by figure," said [C. Glenn] Begley. "I explained that we re-did their experiment 50 times and never got their results." He said they'd done it six times and got this result once, but put it in the paper because it made the best story" (2012).

Goodman noted that the approach taken by Begley and Ellis is a valuable attempt to uncover the truth. He stressed that they did not just do a one-off attempt

[2] The Many Labs Replication Project was a 36-site, 12-country, 6,344-subject effort to try to replicate a variety of important findings in social psychology (Klein et al., 2014; Nosek, 2014; Nauts et al., 2014; Wesselmann et al., 2014; Sinclair et al., 2014; Vermeulen et al., 2014; Müller and Rothermund, 2014; Gibson et al., 2014; Moon and Roeder, 2014; IJzerman et al., 2014; Johnson et al., 2014; Lynott et al., 2014; Žeželj and Jokić, 2014; Blanken et al., 2014; Calin-Jageman and Caldwell, 2014; Brandt et al., 2014). More information is available at Open Science Framework, "Investigating Variation in Replicability: A "Many Labs" Replication Project," last updated September 24, 2015, https://osf.io/wx7ck/.

to replicate and compare the results; they tried every way possible to try to demonstrate the underlying phenomenon, even if that took attempting an experiment 50 times. This goes beyond replication and aims to reveal the true phenomenon; however, he noted that the same language is being used.

There is also no clear consensus about what constitutes reproducibility, according to Goodman. He quoted from a paper by C. Glenn Begley and John Ioannidis:

> This has been highlighted empirically in preclinical research by the inability to replicate the majority of findings presented in high-profile journals. The estimates for irreproducibility based on these empirical observations range from 75% to 90%. . . . There is no clear consensus at to what constitutes a reproducible study. The inherent variability in biological systems means there is no expectation that results will necessarily be precisely replicated. So it is not reasonable to expect that each component of a research report will be replicated in perfect detail. However, it seems completely reasonable that the one or two big ideas or major conclusions that emerge from a scientific report should be validated and withstand close interrogation. (2015)

Goodman observed that in this text the terms "inability to replicate" and "irreproducibility" are used synonymously, and the words "reproducible," "replicable," and "validated" all relate to the process of getting at the *truth* of the claims.

Goodman also mentioned an article from Francis Collins and Larry Tabak (2014) on the importance of reproducibility:

> A complex array of other factors seems to have contributed to the lack of reproducibility. Factors include poor training of researchers in experimental design; increased emphasis on making provocative statements rather than presenting technical details; and publications that do not report basic elements of experimental design. . . . Some irreproducible reports are probably the result of coincidental findings that happened to reach statistical significance, coupled with publication bias. Another pitfall is the overinterpretation of creative "hypothesis-generating" experiments, which are designed to uncover new avenues of inquiry rather than to provide definitive proof for any single question. Still, there remains a troubling frequency of published reports that claim a significant result, but fail to be reproducible.

Goodman summarized the three meanings of "reproducibility" as follows:

1. *Reproducing the processes of investigations*: looking at what a study did and determining if it is clear enough to know how to interpret results, which is particularly relevant in computational domains. These processes include factors such as transparency and reporting, and key design elements (e.g., blinding and control groups).
2. *Reproducing the results of investigations*: finding the same evidence or data, with the same strength.

3. *Reproducing the interpretation of results*: reaching the same conclusions or inferences based on the results. Goodman noted that there are many cases in epidemiology and clinical research in which an investigation leads to a significant finding, but no associated claims or interpretations are stated, other than to assert that the finding is interesting and should be examined further. He does not think this is necessarily a false finding, and it may be a proper conclusion.

Goodman offered a related parsing of the goals of replication from Bayarri and Mayoral (2002):

- *Goal 1: Reduction of random error.* Conclusions are based on a combined analysis of (original and replicated) experiments.
- *Goal 2: Validation (confirmation) of conclusions.* Original conclusions are validated through reproduction of similar conclusions by independent replications.
- *Goal 3: Extension of conclusions.* The extent to which the conclusions are still valid is investigated when slight (or moderate) changes in the co-variables, experimental conditions, and so on, are introduced.
- *Goal 4: Detection of bias.* Bias is suspected of having been introduced in the original experiment. This is an interesting (although not always clearly stated) goal in some replications.

Goodman offered two additional goals that may be achieved through replication:

- *Learning about the robustness of results*: resistance to (minor or moderate) changes in the experimental or analytic procedures and assumptions, and
- *Learning about the generalizability (also known as transportability) of the results*: true findings outside the experimental frame or in a not-yet-tested situation. Goodman added that the border between robustness and generalizability is indistinct because all scientific findings must have some degree of generalizability; otherwise, the findings are not actually science.

Measuring reproducibility is challenging, in Goodman's view, because different conceptions of reproducibility have different measures. However, many measures fall generally into the following four categories: (1) process (including key design elements), (2) replicability, (3) evidence, and (4) truth.

When discussing process, the key question is whether the research adhered to reproducible research-sharing standards. More specifically, did it provide code, data, and metadata (including research plan), and was it registered (for many clinical research designs)? Did it adhere to reporting (transparency) standards (e.g.,

CONSORT,[3] STROBE,[4] QUADAS,[5] QUORUM,[6] CAMARADES,[7] REMARK[8])? Did it adhere to various design and conduct standards, such as validation, selection thresholds, blinding, and controls? Can the work be evaluated using quality scores (e.g., GRADE[9])? Did the researchers assess meta-biases that are hard to detect except through meta-research? Goodman noted that in terms of assessing meta-biases, one of the most important factors is selective reporting in publication. This is captured by multiple terms and often varies by discipline:

- Multiple comparisons (1950s, most statisticians),
- File drawer problem (Rosenthal, 1979),
- Significance questing (Rothman and Boice, 1979),
- Data mining, dredging, torturing (Mills, 1993),
- Data snooping (White, 2000),
- Selective outcome reporting (Chan et al., 2004),
- Bias (Ioannidis, 2005),
- Hidden multiplicity (Berry, 2007),
- Specification searching (Leamer, 1978), and
- p-hacking (Simmons et al., 2011).

Goodman observed that all of these terms refer to essentially the same phenomenon in different fields, which underscores the importance of clarifying the language.

In terms of results, Goodman explained, the methods used to assess replication are not clear. He noted some methods that are agreed upon within different disciplinary cultures include contrasting statistical significance, assessing statistical

[3] The Consolidated Standards of Reporting Trials website is http://www.consort-statement.org, accessed January 6, 2016.

[4] The Strengthening the Reporting of Observational Studies in Epidemiology website is http://www.strobe-statement.org, accessed January 12, 2016.

[5] The Quality Assessment of Diagnostic Accuracy Studies website is http://www.bris.ac.uk/quadas/, accessed January 12, 2016.

[6] See, for example, the Quality of Reporting of Meta-Analyses checklist for K.B. Filion, F.E. Khoury, M. Bielinski, I. Schiller, N. Dendukuri, and J.M. Brophy, "Omega-3 fatty acids in high-risk cardiovascular patients: A meta-analysis of randomized controlled trials," *BMC Cardiovascular Disorders* 10:24, 2010, electronic supplementary material, Additional file 1: QUORUM Checklist, http://www.biomedcentral.com/content/supplementary/1471-2261-10-24-s1.pdf.

[7] The Collaborative Approach to Meta-Analysis and Review of Animal Data from Experimental Studies website is http://www.dcn.ed.ac.uk/camarades/, accessed January 12, 2016.

[8] The REporting recommendations for tumour MARKer prognostic studies website is http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2361579/, accessed January 12, 2016.

[9] The Grading of Recommendations, Assessment, Development and Evaluations website is http://www.gradeworkinggroup.org/, accessed January 12, 2016.

compatibility, improving estimation and precision, and assessing contributors to bias/variability.

Assessing replicability from individual studies can be approached using probability of replication, weight of evidence (e.g., p-values, traditional meta-analysis and estimation, Bayes factors, likelihood curves, and the probability of replication as a substitute evidential measure), probability of truth (e.g., Bayesian posteriors, false discovery rate), and inference/knowledge claims. Goodman noted that there are many ways to determine the probability of replication, many of which include accounting for predictive power and resampling statistics internally.

Goodman discussed some simple ways to connect evidence to truth and to replicability, as shown in Table 3.1. However, he noted that very few scientists are even aware that these measures can show how replication is directly related to strength of evidence and prior probability. A participant questioned shifting the conversation from reproducibility to truth and how to quantify uncertainty in truth. Specifically, since so much of the reproducibility discussion is fundamentally about error, is truth a specific target number in the analysis or is it a construct for which there is a range that would be acceptable? Goodman explained that replicability and reproducibility is a critical operational step on the road to increasing certainty about anything. Ultimately the goal is to accumulate enough certainty to declare something as established. He explained that the reproducibility in and of itself is not the goal; rather, it is the calculus that leads a community to agree after a certain number of experiments or different confirmations of the calculations that the result is true. This is why, according to Goodman, a hierarchy of evidence exists, and the results from an observational study are not treated in quite the same way as a clinical trial even though the numbers might look exactly the same in the standard errors. Some baseline level of uncertainty might be associated with the phenomenon under study, and additional uncertainties arise from measurement issues and covariance. To evaluate the strength of evidence, quantitative and qualitative assessments need to be combined. Goodman noted that some measures directly put a probability statement on truth as opposed to using operational surrogates. Only through that focus can we understand exactly how reproducibility, additional experiments, or increased precision raises the probability that we have a true result. However, he commented that there are no formal equations that incorporate all of the qualitative dimensions of assessing experimental quality and covariant inclusion or exclusion. To some extent, this can be captured through sensitivity analyses, which can be incorporated into that calculus. While the uncertainty can be captured in a Bayesian posterior through Bayesian model averaging, this is not always done and it is unclear if this is even the best way to proceed.

Goodman mentioned a highly discussed editorial (Trafimow and Marks, 2015) in *Basic and Applied Social Psychology* whose authors banned the use of indices related to null hypothesis significance testing procedures in that journal (including

p-values, t-values, F-values, statements about significant difference, and confidence intervals). Goodman concluded that reproducibility does not have a single agreed-upon definition, but there is a widely accepted belief that being able to demonstrate a physical phenomenon in repeated independent contexts is an indispensable and essential component of scientific truth. This requires, at minimum, a complete knowledge of both experimental and analytic procedures (both process-related reproducibility and the strength of evidence). He speculated that reproducibility is the focus instead of truth, per se, because frequentist logic, language, and measures do not provide proper evidential or probabilistic measures of truth. The statistical measures related to reproducibility depend on the criteria by which *truth* or true claims are established in a given discipline. Goodman stated that in some ways, addressing the problems would be a discipline-by-discipline reformation effort where statisticians have a special role.

Goodman reiterated that statisticians are in a position to see the commonalities of problems and solutions across disciplines and to help harmonize and improve the methods that we use to enhance reproducibility on the journey to truth. Statisticians communicate through papers, teaching, and development of curricula, but the curricula to cross many disciplines do not currently exist. He noted that these curricula have to cover the foundation of inference, threats to validity, and the calculus of uncertainty. Statisticians can also develop software that incorporates measures and methods that contribute to the understanding and maximization of reproducibility.

Goodman concluded his presentation by noting that enhancing reproducibility is a problem of collective action that requires journals, funders, and academic communities to develop and enforce the standards. No one entity can do it alone. In response to a participant question about why a journal could not force authors to release their data and methods without the collaboration of other journals, Goodman explained that many journals are in direct competition and no one journal wants to create more impediments for the contributors than necessary. Data sharing is still uncommon in many fields, and a journal may hesitate to require something of authors that other competitive journals do not. However, Goodman noted that many journals are starting to move in this direction. The National Institutes of Health and other funders are beginning to mandate data sharing. The analogous movement toward clinical trial registration worked only when the top journals declared they would not publish clinical trials that were not preregistered. While any one journal could stand up and be the first to similarly require data sharing, Goodman stressed that this is unlikely given the highly competitive publishing environment.

Another participant stated that even if all data were available, the basic questions about what constitutes reproducible research remain because there is not a clear conceptual framework. Goodman observed that there may be more opportunity for agreement about the underlying constructs than about language

because language seems to differ by field. Mapping out the underlying conceptual framework, however, and pointing out that while these words are applied here or there in different ways and cultures, everyone is clear about the underlying issues (e.g., the repeatability of the experiment versus the calculations versus the adjustments). If the broader community can agree on that conceptual understanding, and specific disciplines can map onto that conceptualization, Goodman is optimistic.

A participant questioned the data-sharing obligation in the case of a large data set that a researcher plans to use to produce many papers. Goodman suggested the community envision a new world where researchers mutually benefit from people making their data more available. Many data-sharing disciplines are already realizing this benefit.

Yoav Benjamini, Tel Aviv University

Yoav Benjamini began his presentation by discussing the underlying effort of reproducibility: to discover the truth. From a historical perspective, the main principle protecting scientific discoveries is a constant subjection to scrutiny by other scientists. Replicability became the gold standard of science during the 17th century, as illustrated by the debate associated with the air pump vacuum. That air pump was complicated and expensive to build, with only two existing in England. When the Dutch mathematician and scientist Christiaan Huygens observed unique properties within a vacuum, he traveled to England to demonstrate the phenomenon to scientists such as Robert Boyle (whom Benjamini cited as the first to introduce the methods section into scientific papers) and Thomas Hobbes. Huygens did not believe the phenomenon would be believed unless he could demonstrate it on a vacuum in England.

According to Fisher (1935), Benjamini explained, "We may say that a phenomenon is experimentally demonstrable when we know how to conduct an experiment which will rarely fail to give us statistically significant results." Benjamini noted that this is the p-value, but since the p-value cannot be replicated a threshold is needed. However, there has been little in the literature about quantifying what it means to be replicable (Wolfinger, 2013) until recently with the introduction of the r-value.

Benjamini offered the following definitions to differentiate between reproducibility and replicability:

- *Reproducibility of the study*. Subject the study's raw data to the same analysis used in the study, and arrive at the same outputs and conclusions.
- *Replicability of results*. Replicate the entire study, from enlisting subjects through collecting data, analyze the results in a similar but not necessarily identical way, and obtain results that are essentially the same as those reported in the original study (Peng, 2009; *Nature Neuroscience*, 2013; NSF, 2014).

He noted that this is not merely terminology because reproducibility is a property of a study, and replicability is a property of a result that can be proved only by inspecting other results of similar experiments. Therefore, the reproducibility of a result from a single study can be assured, and improving the statistical analysis can enhance its replicability.

Replicability assessment, Benjamini asserted, requires multiple studies, so it can usually be done within a meta-analysis. However, meta-analysis and replicability assessment are not the same. Meta-analysis answers the following question: Is there evidence for effects in at least one study? In contrast, a quantitative replicability assessment should answer this question: Is there evidence for effect in at least two (or more) studies? This is not in order to buy power but in order to buy a stronger scientific statement.

An assessment establishes replicability in a precise statistical sense, according to Benjamini. If two studies identify the same finding as statistically significant (with p-values P_1 and P_2), replicability is established if the union hypothesis $H_{01} \cup H_{02}$ is rejected in favor of the conjunction of alternatives

$$\text{r-value} = max(P_1; P_2)$$

The r-value is therefore the level of the test according to which the alternate is true in at least two studies. Similar findings in at least two studies are a minimal requirement, but the strength of the replicability argument increases as the number of studies with findings in agreement increases. When screening many potential findings, the r-value is the smallest false discovery rate at which the finding will be among those where the alternate is true in at least two studies.

Benjamini concluded his presentation with several comments:

- Different designs need different methods. For example, an empirical Bayes approach can be used for genome-wide association studies (Heller and Yekutieli, 2014).
- The mixed-model analysis can be used as evidence for replicability.[10] It has the advantage of being useful for enhancing replicability in single-laboratory experiments. The relation between the strength of evidence and power needs to be explored, perhaps through the use of r-values.
- Selective inference and uncertainty need to be introduced into basic statistical education (e.g., the Bonferroni correction method, false discovery rate, selective confidence intervals, post-selection inference, and data splitting).

[10] See Chapter 2 for Benjamini's previous discussion of this topic.

REPRODUCIBILITY AND STATISTICAL SIGNIFICANCE

Dennis Boos, North Carolina State University

Dennis Boos began his workshop presentation by explaining that variability refers to how results differ when an experiment is repeated, which naturally relates to reproducibility of results. The subject of his presentation was the variability of p-values (Boos and Stefanski, 2011); he asserted that p-values are valuable indicators, but their use needs to be recalibrated to account for replication in future experiments.

To Boos, reproducibility can be thought of in terms of two identical experiments producing similar conclusions. Assume summary measures T_1 and T_2 (such as p-values) are continuous and exchangeable.

$$(T_1, T_2) \stackrel{d}{=} (T_2, T_1)$$

then

$$P(T_2 > T_1) = P(T_1 < T_2) = 1/2$$

Thus, before the two experiments are performed, there is a 50 percent chance that $T_2 > T_1$. Boos noted that for p-values, it is not unusual for weak effects that produce a p_1 near, but less than, 0.05 to be followed by p_2 larger than 0.05, which indicates that the first experiment is not reproducible. Boos commented that this is not limited to p-values; regardless of what measure is used, if the results are near the threshold of what is declared significant, there is a good chance that that standard will not be met in the subsequent experiment.

Boos paused to provide some background on standard errors. If a sample mean $\bar{Y}$ is being reported, the standard deviation s divided by the square root of the sample n is usually reported as the standard error, $s/\sqrt{n}$, although this is not always meaningful (Miller, 1986). If regression coefficients $\hat{\beta}$ are reported, the standard error is easy to produce unless model selection is used. However, standard errors are not typically included when using Monte Carlo estimates, p-values, or R^2, all of which can have high variability (Lambert and Hall, 1982; Boos and Stefanski, 2011). He noted that the log scale of p-values is more reliable.

Boos argued that only the orders of magnitude of p-values, as opposed to their exact values, are accurate enough to be reported, with three rough ranges being of interest:

$0.01 < p \leq 0.05$ (often denoted as *)
$0.001 < p \leq 0.01$ (often denoted as **)
$p \leq 0.001$ (often denoted as ***)

He suggested that some alternate methodologies could be useful in reducing variability:

- Bootstrap and jackknife standard errors for log_{10} (p-value), as suggested by Shao and Tu (1996);
- Bootstrap prediction intervals and bounds for p_{new} of a replicated experiment, as suggested by Mojirsheibani and Tibshirani (1996); and
- Estimate of the reproducibility probability P ($p_{new} \leq 0.05$), which assesses the probability that a repeated experiment will produce a statistically significant result (Goodman, 1992; Shao and Chow, 2002).

Using the reproducibility probability, Boos argued that p-values of $p \leq 0.01$ are necessary if the reported results are to an acceptable degree of reproducibility. However, the reproducibility probability has a large variance itself, so the standard error is large. Boos suggested using the reproducibility probability as a calibration but remaining aware of its limitations.

Boos also discussed Bayes calculations in the one-sided normal test (Casella and Berger, 1987). He gave an example of looking at the posterior probability over a class of priors. Using the Bayes factor, which is the posterior probability of the alternative over the null hypothesis, he illustrated that Bayes factors show similar repeatability problems as p-values when the values are near the threshold.

To conclude, Boos summarized that under the null hypothesis, p-values have large variability. The implications for reproducibility (in terms of the reproducibility probabilities $\widehat{RP}$) are as follows:

$$p = 0.01 \quad \widehat{RP} \approx 0.73$$
$$p = 0.001 \quad \widehat{RP} \approx 0.91$$
$$p = 0.0001 \quad \widehat{RP} \approx 0.97$$
$$p = 0.00001 \quad \widehat{RP} \approx 0.99$$

Reporting p-values in terms of ranges (specifically *, **, or ***) may be a reasonable approach for the broader community, he suggested.

Andreas Buja, Wharton School of the University of Pennsylvania

Andreas Buja stated that the research from Boos and Stefanski (2011) made many significant contributions to the scientific community's awareness of sampling variability in p-values and showed that this variability can be quantified. That awareness led to a fundamental question: When seeing a p-value, is it believable that something similar would appear again under replication? Because a positive

answer is not a given, Buja argues that more stringent cutoffs than $p = 0.05$ are important to achieve replicability.

There are two basic pedagogical problems, according to Buja: (1) p-values are random variables; while they appear to be probabilities, they are transformed and inverted test statistics. (2) The p-value random variables exhibit sampling variability, but the depth of this understanding is limited and should be more broadly stated as data set–to–data set variability. Moving from p-value variability to bias, Buja suggested that multiple studies might be more advantageous than a single study with a larger sample because interstudy correlation can be significant.

Buja argued that statistics and economics need to work together to tackle reproducibility. Statistical thinking, he explained, views statistics as a quantitative epistemology and the science that creates protocols for the acquisition of qualified knowledge. The absence of protocols is damaging, and important distinctions are made between replicability and reproducibility in empirical, computational, and statistical respects. Economic thinking situates research within the economic system where incentives must be set right to solve the reproducibility problem. Buja argued that economic incentives (e.g., journals and their policies) should to be approached in conjunction with the statistical protocols.

Journals need to fight the file drawer problem, Buja said, by stopping the chase of "breakthrough science" and publishing, soliciting, and favorably treating replicated results and negative outcomes. He argued that institutions and researchers are lacking suitable incentives. Researchers will self-censor if journals treat replicated results and negative outcomes even slightly less favorably. Also, researchers will lose interest as soon as negative outcomes are apparent. In Buja's view, negative results are important but often are not viewed as such by researchers because publishing them is challenging. Following the example of Young and Karr (2011), Buja argued that journals should accept or reject a paper based on the merit and interest of the research problem, the study design, and the quality of researcher, without knowing the outcomes of the study.

Statistical methods should take into account all data analytic activity, according to Buja. This includes the following:

- Revealing all exploratory data analysis, in particular visualizations;
- Revealing all model searching (e.g., lasso, forward/backward/all-subsets, Bayesian, cross validated, Akaike information criterion, Bayesian information criterion, residual information criterion);
- Revealing all model diagnostics and actions resulting from them; and
- Inferencing steps that attempt to account for all of the above.

In principle, Buja stated that any data-analytic action that could result in a different outcome in another data set should be documented, and the complete set of data-analysis inference should be undertaken in an integrated fashion.

While these objectives are not yet attained, Buja identified some work that is contributing to these goals. Post-selection inference, for example, is a method for statistical inference that is valid after model selection in linear models (Berk et al., 2013). Buja explained that this helps secure against attempts at model selection, including p-hacking. Inference for data visualization is also progressing, according to Buja. In principle, synthetic data can be plotted with and compared to actual data. Sources of synthetic data include permutations for independence tests, parametric bootstrap for model diagnostics, and sampling conditional data distributions given sufficient statistics. If the actual data plot can be differentiated from the synthetic plots, significance can be demonstrated (Buja et al., 2009).

Valen Johnson, Texas A&M University

Valen Johnson began by noting that the nature of p-values does not answer the question of whether a test statistic is bigger than it would be if the experiments were repeated; it does prove whether the null hypothesis is true.

Johnson offered the following review of the Bayesian approach to hypothesis testing: The Bayes theorem provides the posterior odds between two hypotheses after seeing the data. This demonstrates that the posterior probability of each hypothesis is equal to the Bayes factor times the prior odds between the two hypotheses.

$$\frac{P(H_1|x)}{P(H_0|x)} = \frac{f(x|H_1)}{f(x|H_0)} \times \frac{P(H_1)}{P(H_0)}$$

or

$$\text{posterior odds} = \text{Bayes factor} \times \text{prior odds}$$

where the Bayes factor is the integral of the sampling density with respect to the prior distribution under that hypothesis:

$$f(x|H_i) = \int f(x|\theta, H_i)\pi_i(\theta, H_i)\,d\theta$$

and when doing null hypothesis significant testing, the prior distribution for the parameter under the null hypothesis is just a point mass $\pi_0(\theta, H_0)$. Bayesian methods are not used frequently because they make it difficult to specify what the prior distribution of the parameter is under an alternative hypothesis.

Recently there has been some methodology developed called uniformly most powerful Bayesian tests (UMPBTs) that provides a default specification of the prior

$\pi_1(\theta, H_1)$ under the alternative hypothesis (Johnson, 2013a). Johnson explained that the rejection region for UMPBT can often be matched to the rejection regions of classical uniformly most powerful tests. Interestingly, the UMPBT alternative places its prior mass at the boundary of the rejection region, which can be interpreted as the implicit alternative hypothesis that is tested by researchers in the classical paradigm. This methodology provides a direct correspondence among p-values, Bayes factors, and the posterior probability. Johnson noted that the prior distribution under the alternative is selected such that it will maximize the probability that the Bayes factor exceeds the threshold against all other possible prior distributions. He commented that it is surprising that such a prior distribution exists because it has to sustain for every data-generating value of the parameter.

For single-parameter exponential family models, Johnson explained that specifying the significance level of the classical uniformly most powerful test is equivalent to specifying the threshold that a Bayes factor must exceed to reject the null hypothesis in the Bayesian tests. So, the UMPBT has the same rejection region as the classical test. If one assumes that equal prior probability is assigned to the null hypothesis and the alternative hypothesis, then there is equivalence between the p-value and the posterior value and the null hypothesis is true.

Under this assumption and using UMPBT to calculate posterior probabilities and prior odds, a p-value of 0.05 leads to the posterior probability of the null hypothesis of 20 percent. Johnson stressed that the UMPBT was selected to give a high probability to rejecting the null at the threshold so the posterior probability would be greater than 20 percent if another prior were used. Johnson speculated that the widespread acceptance of a p-value of 0.05 reduces reproducibility in scientific studies. To get to a posterior probability that the null hypothesis is true with 0.05, he asserted that one really needs a p-value of about 0.005.

The next question Johnson posed was whether the assumed prior probability of the null hypothesis of only 0.5 is correct. He noted that when researchers do experiments, they do multiple experiments with the same data in multiple testing, conduct subgroup analysis, and have file drawer biases, among other issues. Given all these potential concerns, he speculated that the prior probability that should be assigned to the null should be much greater than 0.5. Assuming a higher prior probability of the null hypothesis would result in an even higher posterior probability. That is, a p-value of 0.05 would be associated with a posterior probability of the null hypothesis that was much greater than 20 percent.

In conclusion, Johnson recommends that minimum p-values of 0.005 and 0.001 should be designated as standards for significant and highly significant results, respectively (Johnson, 2013b).

Panel Discussion

After their presentations, Dennis Boos, Andreas Buja, and Valen Johnson participated in a panel discussion. During this session, two key themes emerged: the value of raising the bar for demonstrating statistical significance (be it through p-values, Bayes factors, or another measure) and the importance of protocols that contribute to the replicability of research.

Statistical Significance

A participant asked how individuals and communities could convince clinical researchers of the benefits of smaller p-values. Johnson agreed that there has been pushback from scientists about the proposal to raise the bar for statistical significance. But he noted that the evidence against a false null hypothesis grows exponentially fast, so even reducing the p-value threshold by a factor of 10 may require only a 60 percent increase in an experiment's sample size. This is in contrast to the argument that massively larger sample sizes would be needed to get smaller p-values. He also noted that modern statistical methodology provides many ways to perform sequential hypothesis tests, and it is not necessary to conduct every test with the full sample size. Instead, Johnson said that a preliminary analysis could be done to see if there appears to be an effect and, if the results look promising, sample size could be increased. Buja commented that, from an economic perspective, fixed thresholds do not work. He argued that there should be a gradual kind of decision making to avoid the intrinsic problems of thresholds.

A participant wondered if the most effective approach might be to try to move people away from p-values to Bayes factors because the Bayes factor of 0.05 is more statistically significant than a p-value of 0.05. Since most researchers are only familiar with the operational meaning of the p-value, the participant argued that keeping the 0.05 value but changing the calculation leading to it may be easier. Johnson noted the challenge of bringing a Bayesian perspective into introductory courses, so educating researchers on Bayes factors is going to be difficult. Because of this challenge, he asserted that changing the p-value threshold to 0.005 might be easier.

A participant proposed an alternative way to measure reproducibility: tracking the ratio of the number of papers that attempted to verify the main conclusion of the report to the number that succeeded (Nicholson and Lazebnik, 2014). Another workshop participant questioned how well such a counting measure would work because many replication studies are underpowered and not truly conclusive on their own. Johnson explained that Bayes factors between experiments multiply together naturally so they serve as an easy way to combine information across multiple experiments.

In response to a participant's question about how much reproducibility might improve if Bayes factors were used rather than p-values, Buja commented that many factors other than statistical methodology could play bigger roles. Boos added that both frequentist and Bayesian analyses will generally come to the same conclusions if done correctly. Johnson agreed that there are many sources of irreproducibility in scientific studies and statistics, and the use of elevated significance thresholds is just one of many factors.

A participant summarized that the level of evidence needs to be much higher than the current standard, be it through Bayes factors or p-values. The participant wondered how to get a higher standard codified in a situation where a variety of models and approaches are being used to analyze a set of data and come up with a conclusion. While a more stringent p-value or higher Bayes factor will help, it may be that neither will provide as much value as single hypothesis tests. But the prior probabilities of various kinds of hypotheses may matter, and that is more difficult to model. Johnson agreed that it is more difficult when working with more complicated models that have high-dimensional parameters. He agreed that the Bayesian principles are in place, but that implementing and specifying the priors is a much more difficult task.

A participant noted that whenever a threshold is used, regardless of what method, approach, process, or philosophy is in place, a result near the threshold would have more uncertainty than a finding within the threshold. Boos agreed that the threshold is just a placeholder and it is up to the community to realize that anything close to it may not reproduce. Johnson disagreed, stating that there needs to be a point at which a journal editor has to decide whether to publish a result or whether a physicist can claim a new discovery. The important takeaway, in Johnson's view, is that this bar needs to be higher.

Protocols

A participant observed that although there are many suggested alternatives to p-values, none of them have been widely accepted by researchers. Johnson responded that UMPBT cannot be computed on all models and that p-values may be a reasonable methodology to use as long as the standards are tightened. Buja reinforced the notion that protocols are important even if they are suboptimal; reporting a suboptimal p-value, in other words, is an important standard.

A participant noted that Buja called for "accounting for all data analytical actions," including all kinds of model selection and tuning; however, the participant would instead encourage scientists to work with an initial data set and figure out how to replicate the study in an interesting way. Buja agreed that replication is the goal, but protocols are essential to make a study fully reproducible by someone else. With good reporting, it may be possible to follow the research path taken by a re-

searcher, but it is difficult to know why particular paths were not taken when those results are not reported. He stressed that there is no substitute for real replication.

ASSESSMENT OF FACTORS AFFECTING REPRODUCIBILITY

Marc Suchard, University of California, Los Angeles

Marc Suchard began his presentation by discussing some recent cases of conflicting observational studies using nationwide electronic health records. The first example given was the case of assessing the exposure to oral bisphosphonates and the risk of esophageal cancer. In this case, Cardwell et al. (2010) and Green et al. (2010) had different analyses of the same data set to show "no significant association" and "a significantly increased risk," respectively. Another example evaluated oral fluoroquinolones and the risk of retinal detachment; Etminan et al. (2012) and Pasternak et al. (2013) used different data sources and analytical methods to infer "a higher risk" and "not associated with increased risk," respectively. Suchard's final example was the case of pioglitazone and bladder cancer, as examined by Wei et al. (2013) and Azoulay et al. (2012), which "[do] not appear to be significantly associated" and "[have] an increased risk," respectively, using the same population but different analytical methods.

Suchard explained that he is interested in what he terms *subjective reproducibility* to gain as much useful information from a reproducibility study as possible by utilizing other available data and methods. In particular, he is interested in quantitatively investigating how research choices in data and analysis might generate an answer that is going to be more reproducible, or more likely to get the same answer, if the same type of experiment were to be repeated.

For example, when examining associations between adverse drug events and different pharmaceutical agents using large-scale observational studies, several studies can be implemented using different methods and design choices to help identify the true operating characteristics of these methods and design choices. This could help guide the choice of what type of design to use in terms of reliability or reproducibility and better establish a ground truth of the positive and negative associations. Another advantage of this approach, Suchard explained, is to better identify the type I error rate and establish the 95 percent confidence interval. He stressed that reproducibility of large-scale observational studies needs to be improved.

As an empirical approach toward improving reproducibility, Suchard and others set up a 3-year study called the Observational Medical Outcomes Partnership (OMOP)[11] to develop an analysis platform for conducting reproducible obser-

[11] The Observational Medical Outcomes Partnership website is http://omop.org, accessed January 12, 2016.

vational studies. The first difficulty with looking at observational data across the world, he explained, is that observational data from electronic health records are stored in many different formats. Addressing this requires a universal set of tools and a common data model that can store all this information, so that everyone is extracting it in the same way when specifying particular inclusion and exclusion criteria. To do this, the data from multiple sources need to be translated into a common data model without losing any information. Then one of many study methods is specified and the associated design decisions are made. Currently, Suchard noted that very few of these choices are documented in the manuscripts that report the work or in the protocols that were set up to initially approve the studies, even though the information might exist somewhere in a database that is specialized for the data researchers were using. If everything is on a common platform, the choices can be streamlined and reported.

Through the OMOP experiment, Suchard and others developed a cross-platform analysis base to run thousands of studies using open-source methods. The results of these studies are fully integrated with vignettes, from the first step of data extraction all the way to publication. He noted that doing this makes everything reproducible both objectively (can the study be reproduced exactly?) and subjectively (can a researcher make minor changes to the study?).

Subject-matter experts were then asked to provide their best understanding of known associations and nonassociations between drugs and adverse events. The experiment now has about 500 such statements about "ground truth," which provide some information on the null distribution of any statistic under any method. Suchard noted that this is important to counter some of the confounders that cannot be controlled in observational studies.

Database heterogeneity can be a problem for reproducible results even when everything else can be held constant, according to Suchard. He explained that looking at different data sets could result in very different answers using the same methodology. Of the first 53 outcomes examined in the OMOP experiment, about half of them had so much heterogeneity that it would not be advisable to combine them, and 20 percent of them had at least one data source with a significant positive effect and one with a significant negative effect.

Suchard then discussed what happens to study results when the data and design are held constant but design choices are varied. He explained that just a small change in a design choice could have a large impact on results.

Another issue is that most observational methods do not have nominal statistical operating characteristics for estimates, according to Suchard. Using the example of point estimates of the relative risk for a large number of negative controls (Ryan et al., 2013), he explained that if the process is modeled correctly and the confidence intervals are constructed using asymptotic normality assumptions, the estimates should have 95 percent coverage over the true value. However, this

proved not to be the case and the estimated coverage was only about 50 percent with both negatively and positively biased estimates. He explained that there needs to be a way to control for the many factors affecting bias so that the point estimates at the end are believable.

When assessing which types of study designs tend to perform best in the real world, Suchard suggested one way to look at the overall performance of the method is to think of it as a binary classifier (Ryan et al., 2013). If a given data source and an open black box analyzer produced either a positive or negative association, the result could be compared with the current set of ground truths and, if these estimates appear to be true, the area under the operator-receiver curve for that binary classification gives the probability of being able to identify a positive result as having a higher value than a negative result. Suchard noted that a group of self-controlled methods that compare the rate of adverse events when an individual is and is not taking a drug tend to perform better in these data with this ground truth than the cohort methods.

Suchard also noted that empirical calibration could help restore interpretation of study findings, specifically by constructing a Bayesian model to estimate what the distribution of the normal values might be using a normal mixture model (Schuemie et al., 2013). The p-value can then be computed as the error under the predicted distribution under the null hypothesis for the observed relative risk. The relative risk versus standard error can then be computed.

To address the findings that observational data are heterogeneous and methods are not reliable (Madigan et al., 2014), the public-private partnership Observational Health Data Sciences and Informatics[12] group was established to construct reproducible tools for large-scale observational studies in health care. This group consists of more than 80 collaborators across 10 countries and has a shared, open data model (tracking more than 600 million people) with open analysis tools.

A participant wondered if medical research could be partitioned into categories, such as randomized clinical trials and observational studies, to assess which is making the most advances in medicine. Suchard pointed out that both randomized control trials and observational studies have limitations; the former is expensive and can put patients at risk, but the latter can provide misinformation. Balancing the two will be important in the future.

Courtney Soderberg, Center for Open Science

Courtney Soderberg began her presentation by giving an overview of the Center for Open Science's recent work to assess the level of reproducibility in

[12] The Observational Health Data Sciences and Informatics website is http://www.ohdsi.org/, accessed January 12, 2016.

experimental research, predict reproducibility rates, and change behaviors to improve reproducibility. The Many Labs project brought together 36 laboratories to perform direct replication experiments of 13 studies. While 10 out of 13 studies were ultimately replicable, Soderberg noted a high level of heterogeneity among the replication attempts. There are different ways to measure reproducibility, she explained, such as through p-value or effect sizes, but if there is a high level of heterogeneity in the chosen measure, it may indicate a problem in the analysis such as a missed variable or an overgeneralization of the theory. This knowledge of the heterogeneity can in itself move science forward and help identify new hypotheses to test. However, Soderberg noted that irreproducibility caused by false positives does not advance science, and the rate of these must be reduced. In addition, research needs to be more transparent so exploratory research is not disguised as confirmatory research. As noted by Simmons et al. (2011), the likelihood of obtaining a false-positive result when a researcher has multiple degrees of freedom can be as high as 60 percent. A large population of researchers admits to engaging in questionable research behavior, but many define it as research degrees of freedom and not as problematic (John et al., 2012).

Soderberg explained that a good way to assess reproducibility rates across disciplines is through a large-scale reproducibility project, such as the Center for Open Science's two reproducibility projects in psychology[13] and cancer biology.[14] The idea behind both reproducibility projects is to better understand reproducibility rates in these two disciplines. In the case of Reproducibility Project: Psychology, 100 direct replications are conducted to determine the reproducibility rate of papers from 2008 and analyze findings to understand what factors can predict whether a study is likely to replicate (Open Science Collaboration, 2015).[15] Moving forward, the project aims to look at initiatives, mandates, and tools to change behaviors. For example, Soderberg explained that open data sharing and study preregistration are not traditional practices in psychology, but the proportion of studies engaging in both has been increasing in recent years. Conducting future reproducibility projects can determine whether behavior change has resulted in higher levels of reproducibility.

John Ioannidis, Stanford University

In his presentation, John Ioannidis noted that a lack of reproducibility could often be attributed to the typical flaws of research:

[13] The Center for Open Science's reproducibility project in psychology website is https://osf.io/ezcuj/wiki/home/, accessed January 12, 2016.

[14] The Center for Open Science's reproducibility project in cancer biology website is https://osf.io/e81xl/wiki/home/, accessed January 12, 2016.

[15] Results reported subsequent to the time of the February 2015 workshop.

- Solo, siloed investigator;
- Suboptimal power due either to using small sample sizes or to looking at small effect sizes;
- Extreme multiplicity, with many factors that can be analyzed but often not taken fully into account;
- Cherry-picking, choosing the best hypothesis, and using post-hoc selections to present a result that is exciting and appears to be significant;
- p-values of $p < 0.05$ being accepted in many fields, even though that threshold is often insufficient;
- Lack of registration;
- Lack of replication; and
- Lack of data sharing.

He explained that empirical studies in fields where replication practices are common suggest that most effects that are initially claimed to be statistically significant turn out to be false positives or substantially exaggerated (Ioannidis et al., 2011; Fanelli and Ioannidis, 2013). Another factor that complicates reproducibility is that large effects, while desired, are not common in repeated experiments (Pereira et al., 2012).

Ioannidis explained that a problem in making discoveries is searching for causality among correlation; the existence of large data sets is making this even more difficult (Patel and Ioannidis, 2014). He proposed that instead of focusing on a measure of statistical inference, such as a p-value, adjusted p-value, normalized p-value, or a base factor equivalent, it may be useful to see where an effect size falls in the distribution of the effect size seen across a field in typical situations.

Another possibility with large databases is to think about how all available data can be analyzed using large consortia and conglomerates of data sets. This would allow for an analysis of effects and correlations to see if a specific association or correlation is more or less than average. However, Ioannidis cautioned that even if a particular correlation is shown, it is difficult to tell if the effect is real. Another approach to the problem is trying to understand the variability of the results researchers want to get when using different analytical approaches (Patel et al., 2015). Ioannidis suggested that instead of focusing on a single result, communities should reach a consensus on which essential choices enter into one analysis and then identify the distribution of results that one could obtain once these analytical choices are taken into account.

He concluded by discussing the potential for changing the paradigms that contribute to irreproducibility problems. He asserted that funding agencies, journals, reviewers, researchers, and others in the community need to identify what goals should be given highest priority (e.g., productivity, quality, reporting quality, reproducibility, sharing data and resources, translational impact of applied research)

and reward those goals appropriately. However, Ioannidis admitted that it is an open question how such goals can be operationalized.

Q&A

A participant congratulated Suchard on his important work in quantifying uncertainty associated with different designs in pharmacoepidemology applications but wondered to what extent the methodology is extendable to other applications, including to studies that are not dependent on databases. Suchard explained that it is difficult to know how to transfer the methodology to studies that are not easy to replicate, but there may be some similarities in applying it to observational studies with respect to comparative effectiveness research. He noted that the framework was not specifically designed for pharmacoepidemology; it was just the first problem to which it was applied.

A participant asked Ioannidis whether—if one accepts that the false discovery rate is 26 percent using a p-value of 0.05—a result with a p-value of 0.05 should be described in papers as "worth another look" instead of "significant"? Ioannidis responded that it would depend on the field and what the performance characteristics of that field would look like, although most fields should be skeptical of a p-value close to 0.05. Recently, he and his colleagues started analyzing data from a national database in Sweden that includes information on 9 million people. Using this database, they are able to get associations with p-values of 10^{-100}; however, he speculates that most are false associations, with such small p-values being possible due to the size of the data set. Similarly for pharmacoepidemology, he stated that almost all drugs could be associated with almost any disease if a large enough data set is used.

A participant asked if there are data about the increase in the number of problematic research papers. Soderberg observed that Simmons et al. (2011) ran simulations looking at how researcher degrees of freedom and flexibility in the analytic choice can cause false positives to go as high as 61 percent. She explained that it is difficult to tell if the real literature has such a high false-positive rate because it is unclear how many degrees of freedom, and how much analytic flexibility, are faced by researchers. However, surveys suggest that it is quite common to make such choices while executing research, and she believes the false-positive rates are higher than the 5 percent that a p-value of 0.05 would suggest. She noted that it is difficult to tell how behavior has evolved over time, so it is hard to measure how widespread reproducibility issues have been in the past.

A participant noted that much of Ioannidis's work is centered on studies with a relatively small sample size, while Suchard's work deals with tens of millions or hundreds of millions of patients, and wondered if there needs to be a fundamental rethinking on how to interpret statistics in the context of data sets where sampling variability is effectively converging to epsilon but bias is still persistent in the

analyses. Ioannidis agreed that with huge sample sizes, power is no longer an issue but bias is a major determinant. He suggested that many potential confounders be examined.

REPRODUCIBILITY FROM THE INFORMATICS PERSPECTIVE

Mark Liberman, University of Pennsylvania

Mark Liberman began his presentation by giving an overview of a replicability experiment using the "common task method," which is a research paradigm from experimental computational science that involves shared training and testing data; a well-defined evaluation metric typically implemented by a computer program managed by some responsible third party; and a variety of techniques to avoid overfitting, including holding test data until after the bulk of the research is done. He explained that the setting for this experiment is an algorithmic analysis of the natural world, which often falls within the discipline of engineering rather than science, although many areas of science have a similar structure.

There are dozens of current examples of the common task method applied in the context of natural language research. Some involve shared task workshops (such as the Conference on Natural Language Learning,[16] Open Keyword Search Evaluation,[17] Open Machine Translation Evaluation,[18] Reconnaissance de Personnes dans les Émissions Audiovisuelles,[19] Speaker Recognition Evaluation,[20] Text REtrieval Conference,[21] TREC Video Retrieval Evaluation,[22] and Text Analysis Conference[23]) where all participants utilize a given data set or evaluation metrics to produce a result and report on it. There are also available data sets, such as the Street View House Numbers,[24] for developing machine-learning and object-recognition algorithms

[16] The Conference on Natural Language Learning website is http://ifarm.nl/signll/conll/, accessed January 12, 2016.

[17] The Open Keyword Search Evaluation website is http://www.nist.gov/itl/iad/mig/openkws.cfm, accessed January 12, 2016.

[18] The Open Machine Translation Evaluation website is http://www.nist.gov/itl/iad/mig/openmt.cfm, accessed January 12, 2016.

[19] The Reconnaissance de Personnes dans les Émissions Audiovisuelles website is https://tel.archives-ouvertes.fr/tel-01114399v1, accessed January 12, 2016.

[20] The Speaker Recognition Evaluation website is http://www.itl.nist.gov/iad/mig/tests/spk/, accessed January 12, 2016.

[21] The Text REtrieval Conference website is http://trec.nist.gov, accessed January 12, 2016.

[22] The TREC Video Retrieval Evaluation website is http://trecvid.nist.gov, accessed January 12, 2016.

[23] The Text Analysis Conference website is http://www.nist.gov/tac/, accessed January 12, 2016.

[24] The Street View House Numbers website is http://ufldl.stanford.edu/housenumbers/, accessed January 12, 2016.

with minimal requirement on data preprocessing and formatting. In the Street View House Numbers case, there has been significant improvement in the performance, especially in terms of reducing the error rate using different methods (Netzer et al., 2011; Goodfellow et al., 2013; Lee et al., 2015).

Liberman pointed out that while the common task method is part of the culture today, it was not so 30 years ago. In the 1960s, he explained, there was strong resistance to such group approaches to natural language research, the implication being that basic scientific work was needed first. For example, Liberman cited *Languages and Machines: Computers in Translation and Linguistics*, which recommended that machine translational funding "should be spent hardheadedly toward important, realistic, and relatively short-range goals" (NRC, 1966). Following the release of that report, U.S. funding for machine translational research decreased essentially to zero for more than 20 years. The committee felt that science should precede engineering in such cases.

Bell Laboratories' John Pierce, the chair of that National Research Council study, added his personal opinions a few years later in a letter to the *Journal of the Acoustical Society of America*:

> A general phonetic typewriter is simply impossible unless the typewriter has an intelligence and a knowledge of language comparable to those of the native speaker of English. . . . Most recognizers behave, not like scientists, but like mad inventors or untrustworthy engineers. The typical recognizer gets into his head that he can solve "the problem." The basis for this is either individual inspiration (the mad "inventor" source of knowledge) or acceptance of untested rules, schemes, or information (the untrustworthy engineer approach) . . . The typical recognizer . . . builds or programs an elaborate system that either does very little or flops in an obscure way. A lot of money and time are spent. No simple, clear, sure knowledge is gained. The work has been an experience, not an experiment. (Pierce, 1969)

While Pierce was not referring to p-values or replicability, Liberman pointed out that he was talking about whether scientific progress was being made toward a reasonable goal.

The Defense Advanced Research Projects Agency (DARPA) conducted its Speech Understanding Research Project from 1972 through 1975, which used classical artificial intelligence to improve speech understanding. According to Liberman, this project was viewed as a failure, and funding ceased after 3 years. After this, there was no U.S. research funding for machine translation or automatic speech recognition until 1986.

Liberman noted that Pierce was not the only person to be skeptical of research and development investment in the area of human language technology. By the 1980s, many informed American research managers were equally skeptical about the prospects, and relatively few companies worked in this area. However,

there were people who thought various sorts of human language technology were needed and that the principle ought to be feasible. In the mid-1980s, according to Liberman, the Department of Defense (DoD) began discussing whether it should restart research in this area.

Charles Wayne, a DoD program manager at the time, proposed to design a speech recognition research program with a well-defined objective evaluation metric that would be applied by a neutral agent, the National Institute of Standards and Technology (NIST), on shared data sets. Liberman explained that this approach would ensure that "simple, clear, sure knowledge" was gained because it would require the participants to reveal their methods at the time the evaluation was revealed. The program was set up accordingly in 1985 (Pallett, 1985).

From this experiment the common task structure was born, in which a detailed evaluation plan is developed in consultation with researchers and published as a first step in the project. Automatic evaluation software (originally written and maintained by NIST but developed more widely now) was published at the start of the project to share data and withhold evaluation test data. Liberman noted that not everyone liked this idea, and many people were skeptical that the research would lead anywhere. Others were disgruntled because the task and evaluation metrics were too explicit, leading some to declare this type of research undignified and untrustworthy.

However, in spite of this resistance and skepticism, Liberman reported that the plan worked. And because funders could measure progress over time, this model for research funding facilitated the resumption of funding for the field. The mechanism also allowed the community to methodically surmount research challenges, in Liberman's view, because the evaluation metrics were automatic and the evaluation code was public. The same researchers who had objected to being tested twice a year began testing themselves hour by hour to try algorithms and see what worked best. He also credited the common task method with creating a new culture where researchers exchanged methods and results on shared data with a common metric. This participation in the culture became so valuable that many research groups joined without funding.

As an example of a project that used the common task method, Liberman described a text retrieval program in 1990 in which four contractors were given money for working on a project. Because of the community engagement across the field, 40 groups from all around the world participated in the evaluation workshop because it offered them an opportunity to access research results for free.

An additional benefit of the common task method is that it creates a positive feedback loop. Liberman explained that when every program has to interpret the same ambiguous evidence, ambiguity resolution becomes a gambling game that rewards the use of statistical methods. That, in turn, led to what has become known as machine learning. Liberman noted that, given the nature of speech and language,

statistical methods need large training sets, which reinforces the value of shared data. He also asserted that researchers appreciated iterative training cycles because they seem to create simple, clear, and sure knowledge, which motivated participation in the common-task culture.

Over the past 30 years, variants of this method have been applied to many other problems, according to Liberman, including machine translation, speaker identification, language identification, parsing, sense disambiguation, information retrieval, information extraction, summarization, question answering, optical character recognition, sentiment analysis, image analysis, and video analysis. Liberman said that the general experience is that error rates decline by a fixed percentage each year to an asymptote, depending on task and data quality. He noted that progress usually comes from many small improvements and that shared data play a crucial role because they can be reused in unexpected ways.

Liberman concluded by stating that while science and engineering cultures vary, sharing data and problems lowers the cost of the data entry, creates intellectual communities, and speeds up replication and extension.

A participant commented that there is a debate about whether subject matter experts can cooperate with computational scientists and statisticians, particularly at this point in time when some statistical methods are outperforming expert rule-based systems. Liberman observed that DARPA essentially forced these communities to work together and managed to change the culture of both fields.

Micah Altman, Massachusetts Institute of Technology

To begin his presentation, Micah Altman surmised how informatics could improve reproducibility. He partitioned these thoughts into two categories: formal properties and systems properties. Formal properties, Altman explained, are properties of information flows and their management that tend to support reproducibility-related inferences. These include transparency, auditability, provenance, fixity, identification, durability, integrity, repeatability, self-documentation, and nonrepudiation. These properties can be applied to different stages and entities, as well as to components of the information system. Systems properties describe how the system interacts with users and what incentives and culture they engender. They include factors such as barriers to entry, ease of use, support for intellectual communities, speed and performance, security, access control, personalization, credit and attribution, and incentives for well-founded trust among actors. There is also the larger question of how the system integrates into the research ecosystem, which includes issues such as sustainability, cost, and the capability for inducing well-founded trust in the system and its outputs.

Altman noted that reproducibility claims are not formulated as direct claims about the world. Instead, they are formulated as questions such as the following:

- What claims about information are implied by reproducibility claims or issues?
- What properties of information and information flow are related to those claims?
- What would possible changes to information processing and flow yield? How much would they cost?

He noted that many elements of the system might be targeted to improve today's situation. People are at the root of the ecosystem, where they affect decisions of theory; select and apply methods; observe and edit data; select, design, and perform analyses; and create and apply documents. The methods also interact, intervene, and simulate in the real world. The researcher's information flow over time ranges among creation/collection, storage/ingest, processing, internal sharing, analysis, external dissemination/publication, reuse (scientific, educational, scientometric, and institutional), and long-term access. Altman reiterated the point made by many of the workshop's speakers that there are many operational reproducibility claims and interventions, such as those outlined in Table 3.2, that are often referred to by the same or inverted names in different communities.

There is an overlap between social science, empirical methods, and statistical computation, according to Altman. In all cases there is a theory guiding the choice of how to conceptualize the world or conceptualize the approach, which leads to algorithms and protocols for solving a particular inference problem. Those protocols or algorithms are implemented in particular coding rules (e.g., instrumentation design and software) and are executed to produce details and context. Informatics tools can be used in a number of environments to deal with information flow, depending on what properties and reproducibility claims are being examined.

TABLE 3.2 Some Types of Reproducibility Issues and Use Cases

Common Labels	Reproducibility Related Issue	Example Interventions
Misconduct, bit rot, author responsibility	Data were fabricated, corrupted, or radically misinterpreted prior to analysis	Discipline/community data archives. NIH genomic data sharing policy
		RetractionWatch; collaborative data collection projects
Misconduct, negligence, confusion, typo, proofreader error, dynamic data problem, versioning problem	Data (referenced by identifier, provided as an instance, described by method) have nontrivial set of semantic differences from that used as input to the publication	Dat, DataHub, DataVerse (versioning)

continued

TABLE 3.2 Continued

Common Labels	Reproducibility Related Issue	Example Interventions
Misconduct, negligence, harmless error	Published analysis algorithm does not correspond to implemented analysis	S/Weave; Compendia; Vistrails
Reproducibility [NSF, 2014; Buckheit and Donoho, 1995]	Variance of estimates given data *instance* and analysis *implementation*	Journal replication data and code archives.
Replicability [King, 1995]		Virtual machine archiving
Replication [NSF, 2014]	Variance of estimates given method *algorithm* and analysis *algorithm*	Protocol archive, *Journal of Visual Experiments*
Reproducibility [King, 1995] independent replication		
Result validation, fact checking	Variance of estimates given *data identifier* and analysis *algorithm*	Data citation standards
Calibration, extension, reuse	Produce new analysis given *data identifier*	Data archives
File drawer problem	Publisher bias toward significant (or expected) results	APS preregistration badge, *Journal of Null Results*
Underreporting (adverse events); data dredging (multiple comparisons)	Author bias toward publishing favored outcomes	Clinical trial preregistration
Data dredging: multiple comparisons; p-hacking	Author bias to creating significant results resulting in difference between stated method/analysis and actual (complete) method/ analysis	Holdout data escrow
Sensitivity, robustness	Variance of support for *claims* across specification change	Sensitivity analysis
Reliability	Variance of support for *claims* across repeated measures, samples	Meta-analysis; Cochrane review
		Data integration
Generalizability	Variance of support for *claims* across different frames	Cochrane review
Laws, truth	Variance of support for *claims* to other populations	Grand challenge?

SOURCE: Courtesy of Micah Altman, Massachusetts Institute of Technology, presentation to the workshop.

To conclude, Altman posed the following questions about how to better support reproducibility with information infrastructure:

- How can we better identify the inferential claims implied by a specific set of (non)reproducibility claims and issues?
- Which information flows and systems are most closely associated with these inferential claims?
- Which properties of information systems support generating these inferential claims?

Altman also thanked his collaborators, namely Kobbi Nissim, Michael Bar-Sinai, Salil Vadhan, Jeff Gill, and Michael P. McDonald, and provided the following references to related work: Allen et al. (2014); Altman and Crosas (2013); Garnett et al. (2013); Altman et al. (2001, 2011); Altman (2002, 2008); Altman and King (2007); Altman et al. (2004); and Altman and McDonald (2001, 2003).

4

The Way Forward: Using Statistics to Improve Reproducibility

The third session of the workshop consisted of three panels discussing how to move forward using statistics to improve reproducibility. The first panel on open problems, needs, and opportunities for methodologic research was moderated by Giovanni Parmigiani (Dana-Farber Cancer Institute and workshop planning committee co-chair) and included Lida Anestidou (National Academies of Sciences, Engineering, and Medicine), Tim Errington (Center for Open Science), Xiaoming Huo (National Science Foundation), and Roger Peng (Johns Hopkins Bloomberg School of Public Health). The second panel on reporting scientific results and sharing scientific study data was moderated by Victoria Stodden (University of Illinois, Urbana-Champaign) and included the following panelists: Keith Baggerly (MD Anderson Cancer Center), Ronald Boisvert (Association for Computing Machinery and National Institute of Standards and Technology), Randy LeVeque (Society for Industrial and Applied Mathematics and University of Washington), and Marcia McNutt (*Science* magazine). The final panel discussion on research as the way forward from the data sciences perspective was moderated by Constantine Gatsonis (Brown University, planning committee co-chair, and chair of the Committee on Applied and Theoretical Statistics) and included Chaitan Baru (National Science Foundation), Philip Bourne (National Institutes of Health), Rafael Irizarry (Harvard University), and Jeff Leek (Johns Hopkins University).

In addition to the references cited in this chapter, the planning committee would like to highlight the following background references: Bossuyt et al. (2003); Couzin-Frankel (2015); Donoho and Huo (2004); Heller et al. (2014); Karr (2014); Laine et al. (2007); Leek and Peng (2015a); LeVeque et al. (2012); Motulsky (2014); Nosek

and Lakens (2014); Political Science Journal Editors (2014); Reiter and Kinney (2011); Stodden (2009a,b); and Stodden et al. (2013b, 2014, 2015).

OPEN PROBLEMS, NEEDS, AND OPPORTUNITIES FOR METHODOLOGIC RESEARCH

Giovanni Parmigiani began the first panel discussion by noting that there should be a more integrated approach to several issues, beginning with terminology. While he commented that many believe the confusion and reversal of terminology across fields is too established to correct at this point, Goodman stated that the underlying conceptual construct behind the terms is shared. Parmigiani sees room to build upon this commonality and identify some construct everyone should refer to in trying to devise reproducibility solutions. He observed that Yoav Benjamini's definitions of and distinction between single-study and multistudy problems reemerges in various attempts at defining terminologies. The distinction between meta-analysis and problems of reproducibility is related to the issue of how to accumulate evidence as it accrues versus how to quantify the extent to which it disagrees. Parmigiani speculated that building on this type of concept is a step that could help the terminology to converge and thus be more useful and conducive to scientific discourse.

He also noted several recurring themes of statistical issues that emerge across fields. He, as well as many other researchers, believes that selection bias is one of the most important of these issues. One aspect of selection bias is hunting for models that provide the desired answer. He suggested a systematic exploration of robustness both in models and experiments as an approach to make headway in this area across fields. He also noted that the ongoing frequentist versus Bayesian debate seems to be dissolving, which may imply that the community is reaching a compromise that could be useful for further progress. A place to start may be an agreement on how to report results (either Bayesian or frequentist) and how to better assess the meaning and significance of a study's results.

Some steps can be taken immediately to identify areas of future work that could benefit multiple fields, according to Parmigiani. He said the statistics community should pay more attention to the issue of reproducibility of prediction across studies, contexts, and data sources. This would allow the scientific community to shift from an abstract definition of *truth* to a paradigm that can be measured more practically and objectively and tied more directly to the decision- and policy-making consequences of studies. For example, pharmacoeconomics, as discussed by Marc Suchard, highlights an arena where it would be possible to have competitions in which research groups, working with given data sets, would be challenged to identify significant associations and interactions, predicting the number of people who have adverse effects over a certain period of time if they take a certain drug.

This exemplifies a prediction question within the context of which reproducibility could be quantified and monitored over time.

Parmigiani noted that this is only one aspect of going beyond statistics' somewhat overused approach of defining tools framed in terms of hypothesis testing. However, he admitted that the community has a lot of work to do to develop tools that are as effective for more complex problems. He also commented that exporting this understanding of reproducibility across sources and subsets of data to the world of Big Data, where data sets are so large that the p-values become meaningless, can be fruitful.

Lida Anestidou, National Academies of Sciences, Engineering, and Medicine

In June 2014, the National Academies of Sciences, Engineering, and Medicine's Institute for Laboratory Animal Research (ILAR), which offers guidance on the use of laboratory animals both in the United States and around the world, convened a workshop on science and welfare in laboratory animal use (NASEM, 2015). Anestidou, who directs ILAR and coordinated that workshop, explained that the purpose of ILAR is to bring the diverse voices in laboratory animal fields together, including members of the public, researchers, veterinarians, laboratory animal facility staff, and committees who have oversight about animal use protocols.

She noted that the use of statistics within the animal research community is diverse and each community member's unique understanding of statistics plays a role in the way reproducibility and other methodologic issues are understood and can be improved. The 2014 workshop discussed fundamental aspects of experimental design of research using animals and animal models, with the goal of improving reproducibility. According to Anestidou, four key themes arose at that workshop:

- Transformation of the research enterprise, specifically systemic issues, scientific training and culture, public perceptions, and incentives for research integrity;
- Interactive assessment of published research;
- Improvement in the reliability of published results; and
- Enhanced understanding of animals and animal models, specifically from clinical research, and proactive planning in preclinical research. This includes reproducibility and the "3Rs" (reduce the number of animals used, refine the methodology, and replace animal models with in vitro and in silico approaches), as well as animal welfare considerations.

An irreproducible study violates the community's notion of ethics and animal welfare, Anestidou explained, because animals are affected and time and money are

wasted. More animals may be needed to repeat a study and animals may not be used appropriately. These are important issues to the laboratory animal community.

She summarized some key points identified by speakers during the ILAR workshop:

- A lack of reproducibility is generally not intentional fraud, and many issues can be linked to flawed experimental design, including statistics and experimental planning.
- C. Glenn Begley defined the following criteria to evaluate journal papers:
  —Is the study blinded? Is a complete set of results shown?
  —Can experiments be replicated?
  —Are positive and negative controls shown? Are statistical tests used (in) appropriately?
  —Are reagents validated?
- Animal models are not poor predictors, and the use of such models does not a priori contribute to reproducibility problems. Rather, speakers at the 2014 workshop identified issues such as small sample sizes, genetic variation among species, and inbred versus outbred strains as leading to reproducibility issues.

Anestidou suggested the following steps to address these issues within the animal models community:

- Educate Institutional Animal Care and Use Committees (IACUCs) so they are able to (re)train investigators on the basics of proper experimental design for animal protocols;
- Design and track metrics of reproducibility involving animal experiments
  —To compare outcomes and trends (i.e., evidence) in association with specific interventions (what about the systemic issues?), and
  —To identify those interventions that appear to be more effective and understand how they may be applied and taught broadly; and
- Energize and interest the broader U.S. research community and involve the laboratory animal veterinary community in the reproducibility conversation.

Tim Errington, Center for Open Science

Tim Errington began by describing some reproducibility issues that arise from researchers' degrees of freedom and explaining how they can essentially short-circuit the scientific process, including a lack of replication (Makel et al., 2012). He also described studies that were designed with low statistical power (Cohen, 1962; Sedlmeier and Gigerenzer, 1989; Bezeau and Graves, 2001); p-hacking (John

et al., 2012); publication bias (Fanelli, 2010); lack of data sharing (Wicherts et al., 2006); and hypothesizing after the results are known (John et al., 2012; Kerr, 1998).

Errington said the research community could take several steps to address these problems. The first is changing the publication process so that review occurs prior to the data collection; this would shift the incentive structure by emphasizing the importance of a study's questions and the quality of its research plans, while lessening pressure to find highly significant or surprising results. The second is study preregistration, which would distinguish between exploratory and confirmatory analysis by requiring information about what data are going to be collected, how the data will be collected, and how the analysis is going to be done. These steps would lead to increased accuracy of reporting, expanded publication of negative results, improved replication research, and enhanced peer review that would focus on the methods and approaches instead of the final result. A handful of journals have already adopted this approach, Errington noted, and the Many Labs and Science Exchange project offers examples of what can be done.[1]

However, Errington explained that adjusting incentives in this way is not enough. He said more tools and technology are needed to couple with the underlying data and methods. Better training is also important, specifically on methodologies that strengthen reproducible statistical analysis and reproducible practices in general, as is increased transparency. In conclusion, he summarized that to improve the scientific ecosystem, technology should *enable* change, training should *enact* change, and incentives should *embrace* change.

Xiaoming Huo, National Science Foundation

Xiaoming Huo began by speaking about the WaveLab project at Stanford University, which began more than 20 years ago and aimed to develop a toolbox to reproduce most of the algorithms available at that time for working with wavelets (Buckheit and Donoho, 1995). He noted that this project showed how important reproducibility is while also providing meaningful workforce development, especially with regard to graduate student training. He emphasized that a focus on reproducibility is a good way to drive stronger methodologic research. For example, he suggested that publications that partially explain how to reproduce software or previously published analysis methods significantly lower the barrier for others to use those methods. However, conducting research into the reproducibility of published work is often viewed as time intensive and outmoded. Because of this, reproducibility work is often not rewarded and may be harmful to those developing academic careers.

[1] The Open Science Framework and the Many Labs and Science Exchange website is https://osf.io/8mpji/, accessed January 12, 2016.

Huo stressed that reproducibility is not only about confirming the work that has been done by someone else; it can also contribute to readability and comprehensibility, especially helping to improve the accessibility of software and methods. Ultimately, Huo explained that the goal of disseminating knowledge is more likely achieved with the use of common terminology.

Huo discussed some National Science Foundation (NSF) programs that help to improve reproducibility. The Advanced Cyberinfrastructure (ACI) Division[2] supports and coordinates the development, acquisition, and provision of state-of-the-art cyberinfrastructure resources, tools, and services essential to the advancement and transformation of science and engineering. In pursuit of this mission, ACI supports a wide range of cyberinfrastructure technologies. In these efforts, ACI collaborates with all NSF Offices and Directorates to develop models, prototypes, and common approaches to cyberinfrastructure.

The Computational and Data-Enabled Science and Engineering program[3] aims to identify and capitalize on opportunities for new computational and data analysis approaches that could enable major scientific and engineering breakthroughs. Research funded under this program relies on the development, adaption, and utilization of one or more of the capabilities offered by advancing research or infrastructure in computation and data, either through cross-cutting or disciplinary programs. Huo noted that the effort's focus on computation and data has a strong connection with reproducibility.

The NSF solicitation for Critical Techniques and Technologies for Advancing Foundations and Applications of Big Data Science and Engineering (BIGDATA) was released in February 2015.[4] According to Huo, this program seeks to fund novel approaches in computer science, statistics, computational science, and mathematics, along with innovative applications in the scientific domain of science, which will enrich the future development of the interdisciplinary field of data science. In conclusion, Huo noted that NSF program officers are always open to hearing new ideas, and he encouraged researchers to reach out directly to discuss potential proposals.

Roger Peng, Johns Hopkins Bloomberg School of Public Health

Roger Peng began by discussing a National Research Council workshop on the Future of Statistical Software (NRC, 1991). At that workshop, Daryl Pregibon set

[2] The NSF's Advanced Cyberinfrastructure Division website is http://www.nsf.gov/div/index.jsp?div=ACI, accessed January 12, 2016.

[3] The NSF's Computational and Data-Enabled Science and Engineering program website is https://www.nsf.gov/funding/pgm_summ.jsp?pims_id=504813, accessed January 12, 2016.

[4] The NSF's BIGDATA website is http://www.nsf.gov/funding/pgm_summ.jsp?pims_id=504767, accessed January 12, 2016.

the stage with the observation that data analysis is a combination of many things that are put together, but that the process as a whole is poorly understood.

Peng noted that he has been writing about reproducibility for about 10 years, and that over this time there has been a tremendous amount of progress in this area, both in the cultures of various communities and in the tools available. For example, he observed that many journals now require data and code(s) when papers are published, and there are entire fields where it has become standard to make code and data available. Tools such as IPython Notebook[5] and Galaxy[6] have been developed to facilitate reproducibility work.

While Peng is encouraged by the developments over recent years, they often do not address the primary hope of researchers that data analysis be more trustworthy and be executed properly. Making data and code available is a good step, according to Peng, because other researchers are then able to correct a broken analysis, but that degree of reproducibility does not prevent incorrect analysis. He said this is analogous to telling someone not to worry if he develops asthma because there are great drugs to control it. What if the asthma could be prevented in the first place?

There are many mistakes in the literature (such as poor experimental design) for which researchers know the solutions, according to Peng. He emphasized the need to better disseminate knowledge that is already available. Many of the reproducibility issues that have arisen over the last couple of years would be best addressed using preventative measures (Leek and Peng, 2015b). In terms of the opportunities for statisticians in particular, Peng suggested that poor data analysis should be proactively prevented, as opposed to something caught after the fact. It is not enough to do peer review or reproducibility work after a bad analysis has been done. Instead, he encourages statisticians to think of the data analysis process more broadly and consider all of it—not just the development of a model—to be a part of statistics. He elaborated that, while experimental design and model development are typically thought of as statistics, the part between has grown massively over the last 10 years. While there is not uniform agreement among statisticians over what part of the process they should be involved in, Peng urges statisticians to be involved in all of it. He asserted that statisticians need to study the process more carefully so they can make recommendations and develop guidelines for how to analyze data appropriately in certain situations, domains, and disciplines. He argued that to take on this new role, the statistics toolbox might need to expand to delve into the realm of experimentation process (i.e., how people do data analysis and what works robustly).

An important step in disseminating information and best practices relates to teaching these techniques and ideas in the simplest possible way. Peng said that

[5] The IPython Notebook website is http://ipython.org/notebook.html, accessed January 12, 2016.
[6] The Galaxy website is https://galaxyproject.org, accessed January 12, 2016.

most researchers need to understand statistics because they are analyzing data, and improving their understanding can help curb some of the poor analyses. He commented that this is a big opportunity for statisticians to embrace the analytic and scientific pipeline and uncover ways to prevent the same reproducibility problems from recurring.

Panel Discussion

A participant noted that in the life sciences, Internal Review Boards (IRBs) in many academic institutions are currently reviewing research proposals before or after funding is received and these IRBs typically have statistical committees. Instead of creating new structures, could IRBs and funding agencies develop a framework for reducing errors in research, which in turn helps increase reproducibility? Errington suggested that having more of the processes tied together would increase the understanding of what the research will be and could lead to improved reproducibility.

Anestidou noted that the animal care and use committees could incorporate statistical subpanels, but there is discussion within the research community about increased regulatory burden, oversight, and paperwork. She wondered how this additional step would fit within the current paradigm. She suggested that the solutions should come from the community instead of being pushed as a top-down regulation.

A participant suggested that government agencies interested in national security analysis know if their analyses are reliable, consistent, and repeatable across analysts. Perhaps reproducibility needs a new framework of analytic engineering to be able to describe how an analysis could be performed and explained to others. Huo noted that in the broader scientific community, the issue of being able to trust an analysis (as Roger Peng described) is important. There are several approaches to this, including increased government regulations or required review, but perhaps a community-based free market model similar to what has been done in some computational communities would be helpful. In such a community, once a method is developed, a paper and the software used are both published. Another approach, Huo noted, is to employ search engines to identify software (using associated comments and reviews) that could potentially be used for a comparison. He suggested that the community could do this sort of work, and those researchers who put more emphasis on reproducibility are more likely to see their papers receive high impact and high citation ratings. He stated that this model is more efficient than having someone else try to impose regulations on the research work.

A participant asked how free statistical consulting relates to the prevention of poor analysis. Errington explained that this type of consulting does not typically carry out a statistical analysis for a researcher; rather, it fosters training and helps

identify methods that would work for a particular research question. He explained that these sorts of interactions aim to help researchers understand that the entire research process, from the way an experiment is designed all the way through to the analysis, is linked and that the entire process needs to be considered holistically. Such services offer advice to help researchers understand what approaches can be used and what resources can help understand the context. The advisors are both in-house and community-driven.

A participant echoed Peng's comment that the scientific community seems not to have absorbed many of the things that statisticians and other designers of research have known for a long time. For example, good scientific practices (such as having a control group or a larger sample size) are well known but not always used. The participant questioned why behavior that researchers know is not optimal is allowed to continue within many communities. The participant then suggested that statisticians play a role here, but they have to partner with people within the disciplines because each culture can only reform itself. The cultures of the disciplines need to change their value systems and understand that every choice made in an analysis is fundamentally an issue of scientific norms and integrity, as opposed to simply moving the dial up or down on the error rate.

Peng agreed that many scientists in many areas know the basics, but his view is that data analysis can quickly get complicated. He also agreed that there is a cultural resistance to accepting this knowledge. However, he suggested that statisticians should bear much of the responsibility to take on this problem and work in the communities to do what is necessary to get them to change. Anestidou agreed that this is an issue of integrity and how to "do" science; she noted that the American Statistical Association has had guidelines about how to "do" statistics for more than a decade. She recalled a statistics professor in her first year of graduate school saying that the methods of statistical analysis need to be chosen before the methodology is set up and work begins. However, she does not see that happening in most cases, which is leading to flawed results. The prevention should start with training because doing better and more reliable analysis is a much larger issue than focusing on the analysis of data that are already collected. Errington commented that a solution to this requires all of the stakeholders to get involved because training and the incentive structure need to be aligned.

A participant noted that the International Initiative for Impact Evaluation[7] funds impact evaluation and systematic reviews that generate high-quality evidence on what works and why. The participant noted that research transparency and better training are almost uniformly agreed upon, but what about all the studies that have already been published? How can incentives truly be changed? How can

[7] The International Initiative for Impact Evaluation website is http://www.3ieimpact.org, accessed January 12, 2016.

researchers be encouraged to do replication studies? Errington agreed that there is value and knowledge to be gained from doing replication research but reiterated that the incentive structure needs to promote it.

A participant suggested that many problems relating to reproducibility could be addressed by adding requirements in the government grant and contracting process. Huo said NSF has been considering how to enhance or impose reproducibility. However, he noted that the funding reality is such that supporting work that attempts to reproduce existing results directly competes with funding other research; this decision, then, has the potential to impact the nation's other science and engineering goals negatively.

A participant then asked how communities might pursue and incentivize improvements to data analysis. Peng said that, at Johns Hopkins, staff members are trying to teach statistics to as many people as possible. And on an individual basis, statisticians there work closely with scientists in their laboratories and environments to improve quality across the board.

A participant stated that prereview of research plans, as would occur with Errington's proposal to revise the publication process, does not allow science to innovate freely. However, the existing IRB and IACUC systems are places where the improvements to analysis could be identified. These committees should include statisticians to evaluate the design of the experiment. The participant also stated that replicability studies might conflict with the 3Rs outlined by Anestidou. Replication is needed as the first step in continued research and should be publishable, but there are ethical concerns of replicating research that may not be of interest to future researchers.

A participant commented that it is encouraging that there is a strong voice emerging from the statistical and data processing community with respect to standards of replicability and improved methodology. However, the funding structure limits what can be accomplished because most investigators are under enormous pressure to get results out quickly in order to demonstrate their productivity and thus qualify for the next increment of funding. There are many instances in which analyses are done prematurely against the advice of statisticians, and researchers shift outcomes or fail to define outcomes adequately at the onset so as to look for an outcome that produces a significant result. The participant noted that it is hard to resist the pressure because statisticians usually work for the investigator and an investigator can look for other statisticians whose recommended adjustments are less burdensome. One possible solution that would address some of these problems is to expand the scope of clinicaltrials.gov, which requires a declaration of the methodology, and to expand this approach to observational studies.

REPORTING SCIENTIFIC RESULTS AND SHARING SCIENTIFIC STUDY DATA

Victoria Stodden began the second panel of the day by noting that the topic of reporting scientific results and sharing scientific data should include sharing scientific code as well. She noted that the topic of dissemination had already come up repeatedly during the workshop, including a discussion of the deliverables that are important when publishing scientific results and the implications this has for dissemination, as well as the methods used to complete studies and report results. She also commented that it is important for the community to consider the public perception of reproducibility.

Keith Baggerly, MD Anderson Cancer Center

Keith Baggerly explained that he has been associated with reproducibility efforts for a few years, motivated by a number of cases where he encountered process failures. This, at times, has led him to explore the development of forensic bioinformatics, which is the art of taking reported results and raw data and inferring what methods were used. He commented that, while this is a useful and often informative art, it does not scale and cannot be used system-wide.

He offered the following as a summary of major takeaway messages he had heard in the workshop:

- The statistical community needs to figure out specific steps that it could contribute to the reproducibility effort.
- The strength of the evidence for a claim presented goes from (a) same results from same source data to (b) same results from new data to (c) aggregate results from lots of data. Baggerly noted that the latter is the goal but that merely getting the same results from the same data can be immensely complicated.
- Research communities need a clearer understanding of the significance cutoff that is acceptable.

Baggerly noted that the case studies discussed during the workshop highlight some notable issues affecting reproducibility, particularly the complications in drawing inferences from large-scale data sets. When utilizing large-scale data sets, Baggerly warns that it is important not to focus on small variation that can be caused by the batch effects that are present, and he reminded the audience that Benjamini had discussed some ways to account for this. A related idea is to look for big effects, particularly in genomics (Zilliox and Irizarry, 2007). Large databases allow for the scale of data noise to be estimated, and that can be used to identify large effects.

Baggerly commented that the community is fortunate to have these large data sets in part because they make it easier to identify real results. In the case of the Cancer Genome Atlas,[8] he explained that the 10,000 samples across 30 different tissues could be viewed as different replications with different disease types, which can help identify where defects are or which tissues have extremely high expression of a gene. He cautioned that it is important not to focus solely on p-values for large data sets, because those values tend to be small. Rather, he suggested that the effect size also be quantified to see if it is big enough to be of practical relevance. Baggerly cautioned that there are still flaws in data processing that come up in the case examples, and these highlighted some of the reasons statisticians need to be involved.

He explained that his main recommendation in terms of reporting is to include sanity checks because multidisciplinary big data studies magnify the chances for inadvertent errors. He elaborated that there are some ways to avoid this, in part through the process of pointing out this possibility to collaborators and soliciting their explanations, plotting by run date, and prespecifying positive and negative controls. As an example, he encourages genomics researchers to write down, before analyzing their data, a short list of the genes that they expect will be changed in response to treatment and the directions in which they should change. This step forces researchers to think about what the results should be before the analysis is performed, thus giving both the analysis team and the data suppliers a way of checking and calibrating results. The positive and negative controls come down to considering (1) what should be seen after the analysis, and (2) what results would indicate that the treatment resulted in no significant differences.

Ronald Boisvert, Association for Computing Machinery and National Institute of Standards and Technology

Ronald Boisvert discussed some of the efforts he has been involved with in the course of his position as a member and former co-chair of the Association for Computing Machinery (ACM) Publications Board, where issues related to reproducibility and data sharing are currently being considered. The ACM is the world's largest scientific and educational society in computing with a substantial publication program (45 research journals, 8 magazines, 26 newsletters, and approximately 450 annual conference proceedings) and an extensive digital library (with more than 430,000 full-text articles and more than 2,300,000 bibliographic records covering the entire computing field).

He explained that publishers could support reproducibility through journal policy mandates as well as by establishing incentives that encourage investigators

[8] The Cancer Genome Atlas website is http://cancergenome.nih.gov, accessed January 12, 2016.

to give greater attention to reproducibility. Publishers can also provide platforms for archiving supplementary material such as data and codes.

Boisvert commented that, on the surface, reproducibility in computing research should be easier than other areas of science because studies are typically carried out computationally, and computational experiments are more easily portable than physical experiments. He noted, though, that this is less so when the research is in areas such as hardware and human-computer interaction.

An early success for the computing community in evaluation and distribution of research software was the *ACM Transactions on Mathematical Software (TOMS)*, which publishes research on the implementation of algorithms for solving standard mathematical problems such as systems of linear equations and partial differential equations. These implementations are packaged into reusable software bundles, which are refereed at the same time that the text of the paper is refereed. The referee gets the software, tries to run it, inspects it, and decides whether it is a useful contribution. Evaluation criteria include aspects such as code structure, usability, documentation, efficiency, and portability. *TOMS* has published more than 450 such papers since 1975, representing about one-third of the papers published in the journal. The software is made available in the ACM Digital Library as supplementary material associated with the paper. The capability of archiving such supplementary material has been available to all ACM publications since 1998, although it is not well promoted and the uptake of data is relatively small. Nevertheless, within the smaller mathematical software community, the desire of researchers to have their code used by others, along with the seal of approval coming from the refereeing process, has sustained the flow of ACM's "Collected Algorithms" for 40 years.

Boisvert noted that other ACM journals have tried without success to replicate what *TOMS* has done. These other journals include the *Journal of Experimental Algorithms*, which has since morphed into a traditional publication, and the *Journal of Educational Resources in Computing*, which is now defunct.

ACM currently encourages pilot efforts to strengthen the reproducibility of papers in its journals and conferences. For example, since 2008 the Special Interest Group on the Management of Data's (SIGMOD's) main conference for database research has had a voluntary process for accepted papers to undergo reproducibility reviews by a committee. Authors submit the software and execution instructions, and these materials are judged on criteria such as sharability, coverage, and flexibility. Papers that pass the review get a "reproducible label"[9] to indicate that the paper was carefully done in a certain sense. Over the years, the standards and procedures for doing the review and the terminologies have changed as the com-

[9] The SIGMOD Reproducibility website is http://db-reproducibility.seas.harvard.edu, accessed January 12, 2016.

munity has gained experience in this practice. The acceptance of this certification process has been fairly high. For example, 35 of the 88 papers accepted in 2011 participated and 24 were confirmed "repeatable" based on a number of criteria (Freire et al., 2012).

Within the programming languages and software engineering community, the issue of reproducibility has been taken on directly, according to Boisvert. There are 11 major conferences on programming language and software engineering that carry out a process known as artifact evaluation,[10] with more than 13 conference sessions participating since 2011. He noted that the optimal committee-based evaluation process for accepted papers has two evaluators per artifact, typically graduate students and postdoctoral fellows. The criteria are similar to those used by SIGMOD in that they look at the packaging, reproducibility, implementation, and usability. This step is beginning to take off in the community: for one particular conference in 2014, 20 out of 52 accepted papers volunteered for this evaluation and 12 passed. While not a requirement, in many cases the artifacts are subsequently made available for download.

The *ACM Transactions on Mathematical Software* recently extended its replicability review process to the two-thirds of its papers in which software is not submitted for review and distribution.[11] Papers in this journal typically present new algorithms and compare them to existing methods via some form of computational experiment, according to Boisvert. Authors of papers that have been accepted subject to minor revisions can opt for an additional "replicability review." In that process, a single reviewer works collaboratively with the authors to replicate the computational results that contribute to the main conclusions of the paper. The reviewer then writes a short paper on the experience, which is published along with the original paper. Authors are incentivized by having the label "Replicated Computational Result" affixed to their papers, while reviewers are incentivized by having a publication of their own.

Boisvert emphasized that while the ACM Publications Board would like to propagate these practices throughout all of its publications, the society understands that success depends on subcommunity acceptance. And before each subcommunity can develop its own procedures for the review process, uniform terminology tied to baseline review standards needs to be developed in order to enable meaningful labeling of papers that have undergone some form of replicability review.

[10] The Artifact Evaluation website is http://www.artifact-eval.org, accessed January 12, 2016.

[11] ACM Digital Library, "Editorial: ACM TOMS Replicated Computational Results Initiative," http://dl.acm.org/citation.cfm?doid=2786970.2743015, accessed January 12, 2016.

Randy LeVeque, Society for Industrial and Applied Mathematics and University of Washington

Randy LeVeque provided some examples of reproducibility work he has been doing over the last 20 years. For example, he co-developed an open-source software package for solving wave propagation problems through his numerical analysis and scientific computing research, studied software applications such as tsunami modeling and hazard assessment, and—as chair of the Society for Industrial and Applied Mathematics (SIAM) Journals Committee—advised the SIAM vice president for publications on issues related to journals. LeVeque also discussed his involvement with the eScience Institute at the University of Washington through the Reproducibility and Open Science Working Group, which aims to change the culture and the way data science is done. This group has regular monthly seminars on reproducibility and other open science issues. This effort began as a single-campus reproducibility effort, where researchers could submit something that they planned to publish and ask other people to download it, run the code, and evaluate the clarity of the instructions. The current goal is to increase the scale of this resource.

The SIAM Journal program has 15 high-impact research journals in applied mathematics. Traditionally, LeVeque explained, supplementary materials had not been published with these journals, but beginning in 2013, editorial boards could determine whether or not they wanted to support additional materials. He emphasized that the idea of supplementary materials was foreign to many researchers in applied mathematics. For example, of the approximately 1,500 articles SIAM publishes each year, only 38 articles have published unrefereed supplementary materials. Two SIAM journals have a longer history of having refereed materials associated with them: the *Journal of Dynamical Systems* uses the DSWeb[12] and the *Journal on Imaging Science* partners with Image Processing On Line.[13]

Several other SIAM journals focus on publishing software,[14] and LeVeque is interested in ensuring people get credit for working on software because it often requires a large investment of time and represents an encapsulation of algorithms and knowledge. He observed, though, that publishing research code for processing, analyzing, and visualizing data; for testing new methods or algorithms; and for computational science or engineering problems is rare in applied mathematics and computation science and engineering.

[12] The DSWeb website is http://www.dynamicalsystems.org/hp/hp/, accessed January 12, 2016.

[13] The Image Processing On Line website is http://www.ipol.im, accessed January 12, 2016.

[14] Other journals focused on publishing software include the *ACM Transactions on Mathematical Software, SIAM Journal on Scientific Computing (Software Section), Journal of Open Research Software, Open Research Computation, Journal of Statistical Software, Geoscientific Model Development,* and *PeerJ Computer Science,* among others.

Making supplementary materials more widely available, both online as well as permanently archived, would advance the field, according to LeVeque. He mentioned two sites, Zenodo[15] and Figshare,[16] that archive material, assign it a digital object identifier, and allow it to link automatically to GitHub.[17] LeVeque recommends that researchers be encouraged to use one of these options for sharing the code that goes along with their research papers.

The culture needs to change, according to LeVeque, and there are still questions about incentives versus requirements. From his perspective, journal publications continue to be valued more highly than software and data sets, and writing another paper is rewarded more than making existing code more reliable and sharable. In LeVeque's view, code and data need to be "first-class objects" in research; he suggested that the broader scientific community needs to imagine a new world in which all data is freely available online and the incentive to hoard data is eliminated.

LeVeque noted that institutional roles concerning code and data sharing are important, particularly regarding whether the institution or the researcher owns the software developed at a given university and what that means for making code open source. He also commented that some curricular changes are needed in computational science, starting at a very basic level with early programming, statistics, and numerical analysis courses. He argued that topics such as version control, code review, and general data hygiene (such as management, metadata, and posting) should be taught early.

He commented that computational mathematicians write papers about numerical methods, often containing a new algorithm, and spend weeks cleaning up the theorems in the paper, but they do not want to spend any time at all cleaning up the code and making it available to others. Traditional mathematics does not struggle as much with reproducibility, according to LeVeque, because proofs are required to publish a theorem. According to David Hume (1738): "There is no . . . mathematician so expert in his sciences, as to place entire confidence in any truth immediately upon his discovery of it. . . . Every time he runs over his proofs, his confidence increases; but still more by the approbation of his friends; and is raised to its utmost perfection by the universal assent and applauses of the learned world." LeVeque argued that computational mathematics should embrace this approach more because it is difficult to evaluate the accuracy of the programs if they have not been cleaned up, published, and peer reviewed.

In conclusion, he noted that many of the arguments against publishing code seem ludicrous when applied to proofs (LeVeque, 2013):

[15] The Zenodo website is http://zenodo.org, accessed January 12, 2016.
[16] The Figshare website is http://figshare.com, accessed January 12, 2016.
[17] The GitHub website is https://github.com, accessed January 12, 2016.

- The proof is too ugly to show anyone else.
- I didn't work out all the details.
- I didn't actually prove the theorem—my student did.
- Giving the proof to my competitors would be unfair to me.
- The proof is valuable intellectual property.
- Including proofs would make math papers much longer.
- Referees would never agree to check proofs.
- The proof uses sophisticated mathematical machinery that most readers/referees don't know.
- My proof invokes other theorems with unpublished (proprietary) proofs. So it won't help to publish my proof—readers still will not be able to fully verify its correctness.
- Readers who have access to my proof will want user support.

He hopes that 300 years from now, people will look back and see this as a transition time when science moved into doing things very differently.

Marcia McNutt, Science Magazine

Marcia McNutt began by noting that the scientific community is embracing the concept of reproducibility quickly. She believes that, in the future, the past couple of years will be identified as the period of time in which the community (i.e., funding agencies, journals, universities, and researchers) recognized it had to take reproducibility seriously and come up with better practices and solutions across all disciplines for the sake of the reputation and quality of science.

She noted that the spectrum of reproducibility (Ioannidis and Khoury, 2011) includes the low end (minimum standard) of repeatability—where another group can access the data, analyze it using the same methodology, and obtain the same result—and the high end (gold standard) of replication—where the study is repeated start to finish, including new data collection and analysis, using fresh materials and reagents, and obtains the same result. For some fields of science, McNutt noted that true replication is not possible. For example, an earthquake cannot be repeated, and forests evolve, so whenever time is a vector in an analysis, exact replication is impossible and the best that can be done is to take the data and analyze them again. There is a certain degree to which you might be able to repeat or generate new data, but it is never going to be an exact repeat.

McNutt explained that the approach at *Science* has been to acknowledge that the differences in fields and communities lead to different reproducibility issues. The journal, then, needs to work with these fields and communities to find the best practices, procedures, and policies that raise the standards for transparency and promote reproducibility. *Science* started with the assumption that a study's

reproducibility or lack of reproducibility does not necessarily mean it is right or wrong, respectively.

She explained that there are many examples of this, including one in which three top laboratories took a global data set of earthquake sources and receivers and analyzed the data to show that there were bumps on the boundary between the Earth's core and mantle. This very reproducible result was widely shown on covers of famous journals and was a source for speculation on the creation of Earth's geodynamo and the coupling mechanism between the differential rotations of the Earth's core and mantle. McNutt emphasized that this was a very solid, reproducible result with major geodynamic repercussions for how Earth behaves. However, the result was fundamentally wrong. All the analyses were preconditioned on the earthquake sources being located in the major subduction zones and the earthquake seismometers being located on the continents. This bias in source receiver locations, when put in a spherical harmonic representation, led to the artifact of bumps on the core/mantle boundary.

Following a workshop on the topic of promoting reproducibility in the preclinical sciences, *Science* published an editorial recommending best practices for transparency in those sciences (McNutt, 2014). That editorial, signed by representatives of 120 journals, recommended that researchers discuss the following information in order to publish:

- Power analysis for how many samples are required to resolve the identified effect,
- Random assignment of samples to treatment and controls,
- Blinding of experimenter to which samples were in the treatment and which were in the controls, and
- Data availability.

A goal was that improved transparency of these four experimental protocols would allow reviewers and readers to gain a level of confidence in the results. McNutt noted that authors are not required to follow these protocols; they are only required to state whether or not they did so.

A follow-on workshop focused on the social and behavioral sciences and resulted in a general document that could be applied more broadly beyond those fields, McNutt stated. That document includes a number of guidelines (Nosek et al., 2015) that journals can choose to follow:

- Tier 1: Asking author to declare what was done,
- Tier 2: Conforming to a community standard, or
- Tier 3: Verifying that the standard was followed.

An upcoming third workshop focusing largely on the availability of data and sample metadata as they pertain to reproducibility in the field sciences will include representatives from journals, data repositories, funders, and the scientific community. A fourth workshop is being planned to focus on the computational sciences.

In conclusion, McNutt noted that *Science* added several statisticians to its board of reviewing editors to help screen and identify papers that may need extra scrutiny for the use of statistics or numerical analysis. She said this addition has raised the journal's standards.

Panel Discussion

Victoria Stodden began the panel discussion by asking each of the panelists to give one or two concrete recommendations for improving how science is conducted, reported, disseminated, or viewed by the public. The following recommendations were offered:

- *Establish publication requirements for open data and code.* Journal editors and referees should confirm that data and code are linked and accessible before a paper is published. (Keith Baggerly)
- *Clarify strength of evidence for findings.* The strength of evidence should be clearly stated for theories and results (in publications, press releases, etc.) to ensure that initial explorations are not misrepresented as being more conclusive than they actually are. (Keith Baggerly)
- *Align incentives.* Communities need to examine how to build a culture that rewards researchers who put effort into verifying their own results rather than quickly rushing to publication. (Marcia McNutt)
- *Improve training.*
  - —Institutions need to make extra efforts to instill students with an ethos of care and reproducibility. (Marcia McNutt)
  - —Universities need to change the curriculum to incorporate topics such as version control, code review, and general data management, and communities need to revise their incentives to improve the chances of reproducible, trustworthy research in the future. Steps to improve the future workforce are necessary to keep the public trust of science. (Randy LeVeque)
  - —Many graduates are well steeped in open-source software norms and ethics, and they are used to this as a normal way of operating. However, they come into a scientific research setting where codes are not shared, transparent, or open; instead, codes are being built or constructed in a way that feels haphazard to them. This training disconnect can interfere with mentorship and with their continuation in science. Better under-

standing of these norms is needed in all levels of research. (Victoria Stodden)

—Prevention and motivation need to be components of instilling the proper ethos. This could be part of National Institutes of Health (NIH)-mandated ethics courses. (Keith Baggerly)

- *Clarify terminology*. A clearer set of terms is needed, especially for teaching students and creating guidelines and best practices. Some examples of how to do this can be found within the uncertainty quantification community, which successfully clarified the terms *verification* and *validation* that were almost used synonymously 10-15 years ago. (Ronald Boisvert)

Regarding the problem of rushing research into publication without consideration of the effects if it is not replicable, a participant stated that there should be mechanisms available to make replications (both positive and negative) better known. McNutt noted that eLife, for biomedical sciences, and the Center for Open Sciences, for the social sciences, are already making such efforts.

A participant noted that many of the journal-sponsored workshops on reproducibility focus on operational issues, such as having transparency, making data available, cataloging, and developing computing infrastructure that allow for the data to become available. That focus overlooks other critical questions: What constitutes evidence of reproducibility (which requires a conceptual framework)? How is reproducibility defined? Who decides whether something is reproducible? How can reproducibility be assessed on an evidentiary basis? The participant wondered how the current machinery is helpful in those efforts and how to make it more clear to researchers what they and the community should be looking for when checking for reproducibility. The participant stated that more development on the conceptual and evidence-base level is needed. LeVeque added that there is still a lot of uncertainty about what exactly should be expected in computational reproducibility even with respect to terminology; for example, there is not a uniform agreement on the terms reproducible, repeatable, and replicable. He noted that the first step in defining an evidence base is having a clear understanding of the terms.

Another participant suggested that one of the ways to change incentives is to make replication research more broadly publishable, possibly through the use of short replication study papers. Boisvert agreed that allowing these replication papers to be published is important for the incentive structure. He referred back to his discussion of how the *ACM Transactions on Mathematical Software* checks for reproducibility as part of the review process, which in part allows reviewers to publish their experiences in doing the replication. Boisvert pointed out that the journal does not currently have a policy of what to do if the replication work fails to reproduce the study findings. Baggerly and a participant agreed that many

organizations and journals are struggling with the same question of how to handle irreproducibility.

A participant commented that researchers at Stanford are beginning to develop a short online module on proper training regarding issues of reproducibility, which is intended to be added to the responsible conduct of research course that students are required to take.

A participant asked how science journalists can help advance reproducibility itself, as well as the public's perception of reproducibility. McNutt suggested that journalists avoid overstating the results of a study (which is a common problem with university press releases) and clarify the caveats and limitations associated with new findings, as well as what might have led the new work to a different conclusion than had previously existed. Baggerly suggested following up on novel findings from the recent past, assessing their implications, and evaluating how well the findings have held up. Stodden commented that journalists can be somewhat hesitant about interacting with the scientists after a story is drafted. Acknowledging the time constraints, she suggested it would be helpful if there were a consistent ethos about having scientists sign off on all quotes. The participant noted that some of the publications she has worked for ask reporters to look through the original papers to assess if the data analysis looks like it was done appropriately. The publications do not want to report on results that end up being undermined by bad data analysis and later criticized on statistics blogs. She agreed that a better system with more open lines of communications is needed.

A participant commented that sensitivity analysis is essential to reproducibility in that the analytical methods used must be assessed to see which are the most sensitive to noise in the study process and how to make them less sensitive. This is at the core of figuring out why something is not reproducible. He wondered if there is a way to get a better scientific infrastructure beyond just journal publishing such as active, open-use databases. Such a system would foster a good interchange among disciplines with regard to how results are reported in various disciplines so that best practices can be accepted and improved upon by other fields. Boisvert noted that uncertainty quantification in computational science, which is related to sensitivity analysis, is a very important consideration to which few people dedicate time. Within the applied mathematics community, there is a large effort in understanding how to do uncertainty quantification of models and simulations, which leads directly to understanding whether results are reproducible. Baggerly agreed that more sensitivity analyses would be helpful and that there needs to be better training in this area. He noted that assessing variation in larger databases is a form of sensitivity analysis and may be about as good as can be done in those cases. McNutt noted that there is new laboratory software entering beta testing that can track laboratory results to reveal systemic issues, such as equipment degradation, and help identify sources of bias and error in results (Gardner, 2014). The participant suggested that the community

could evolve with the help of a standards development organization that conducts interlaboratory studies and makes results available to the public so other researchers can identify similar sensitivities in their own laboratories. Baggerly agreed that having a built-in system to spot interlaboratory variations would be ideal.

THE WAY FORWARD FROM THE DATA SCIENCES PERSPECTIVE: RESEARCH

Constantine Gatsonis opened the final workshop panel by highlighting some of the previously identified themes. The first relates to statistical thinking and determining evidence of reproducibility. He echoed previous speakers in stating that the discussion of reproducibility is at a critical point and the issue is well recognized across the scientific and policy domains. In some specific areas, approaches toward assessing evidence for reproducibility have been developed that are applicable to a particular area of research. However, he emphasized that there is not a broadly accepted framework for conceptualizing and assessing reproducibility. Some key questions that need to be addressed in such a framework include the following, according to Gatsonis:

- What is meant by *reproducibility*?
- What evidence is needed to support reproducibility?
- How should experiments be designed?
- What is the role of publishing in supporting that enterprise?
- How stringent should the evidence be before a result is declared reliable from a reproducibility perspective?
- What is the right p-value or Bayes factor?

Gatsonis stressed that these are important issues about evidence that highlight the lack of consensus among scientific communities. Individual scientific communities are developing solutions to portions of the challenge, and certain areas are evolving quickly, such as policy, computing approaches, and IT. However, a more conceptual framework still needs to be developed. The National Academies' Committee on Applied and Theoretical Statistics, which organized this workshop, is looking for ways to move this forward.

Another open issue identified by Gatsonis is what researchers and students should be taught about reproducibility. He emphasized that there needs to be explicit curricula with courses that address reproducibility directly. However, he argued that before anyone could develop this curriculum, a broadly accepted framework for what constitutes strong evidence for reproducibility is needed. He noted that different scientific communities are at the stage of developing structures and processes, but that the basic scientific consensus is still evolving.

Philip Bourne, National Institutes of Health

Philip Bourne explained that he would build on Lawrence Tabak's ideas from earlier in the workshop and go into more detail with respect to data. He observed that NIH's data science strategy incorporates statistical rigor, replication, and reproducibility by focusing on community, policy, and infrastructure, most of which is being done through the Big Data to Knowledge (BD2K) initiative.

Bourne noted a few key issues, the first of which is the significant time required to reproduce research (Garijo et al., 2013). The second is the insufficient reporting of data and a lack of negative data. He suggested that the way to address this gap is through the use of a Commons, which is essentially a shared space where research objects are posted (e.g., data, software associated with analyzing that data, statistical analysis, narratives, and final publications). A third issue is p-hacking and robust research training revolving around the best use of statistics and analytics. NIH is creating a training coordination center to begin collecting and recording courses offered and materials available (both physical and virtual), as well as cataloging the analysis training landscape and identifying gaps which might require additional funding.

Incentives, Bourne asserted, are a key aspect of encouraging reproducibility and statistical rigor in research. NIH's policies are changing with respect to data sharing. Currently, the Office of Science and Technology Policy directs that any grant over $500,000 must have a data-sharing plan, but soon this requirement will be extended to all grants. Bourne commented that while some of the incentives come from funders, many come from the community. For example, he does not believe data are regarded highly enough in the realms of scholarship. Perhaps endorsing data citations in new ways would be helpful, as is being done through the National Library of Medicine's PubMed and PubMed Central.

Chaitan Baru, National Science Foundation

Chaitan Baru affirmed that the issue of data is important to NSF; the effectiveness of their current data management plan will be evaluated over time. NSF funds individual research proposals in this area, as well as a community group that studies ethics concerns. Baru explained that there are three primary areas that make up data science: computer science, statistics, and ethics and social issues. He discussed the 2014 Big Data Strategic Initiative[18] workshop that brought together federal agencies, academia, and industry to discuss agencies' strategies for dealing with data and data analysis. An important theme that emerged from that workshop

[18] The NSF's Big Data Strategic Initiative Workshop website is http://workshops.cs.georgetown.edu/BDSI-2015/, accessed January 12, 2016.

was education and training for current researchers and for the next generation. As a result, his office intends to run another workshop on how data science curricula should be designed. He emphasized that the concepts of reproducibility and repeatability would be essential elements within a data science curriculum.

Baru concluded by recognizing the ACM database conference, which instituted the notion of looking at the repeatability of results in the papers submitted, and stressing that more such work would help the community. He commented that it is difficult for a funding agency to advance a cultural change that is not already occurring in a community. If the norm develops, then the internal pressure and behaviors such as control begin to move in the right direction. For example, as a community, ACM SIGMOD generated the notion of a test-of-time award, in which conference proceedings and papers from 10 years ago are evaluated and an award is given to the paper that had the most impact. A participant later commented that there are branches of ethnography that study how people collaborate with different branches of science and how people and cultures change; such work may provide some insight into how to change aspects of the scientific cultures relating to reproducibility.

Rafael Irizarry, Harvard University

Rafael Irizarry echoed Philip Bourne's message about the importance of education: better training is the best way to prevent errors in methods and analysis and thus improve reproducibility. He elaborated that improved education is particularly important now as many fields are transitioning from a data-poor to a data-driven state, and many researchers are becoming data analysts out of necessity. He commented that while he is not an expert in reproducibility, he has been working in biomedical data science for 20 years, helping to manage data and make discoveries. During this time, biomedical sciences have become data intensive and many researchers must now be proficient in data management, data wrangling, computer algorithm optimization, and software development to implement methods. Irizarry noted that a relatively small investment of time and resources at the beginning of a project has the potential to improve reproducibility and save a lot of time in the end.

He highlighted a few readily available tools created by data analysts to improve reproducibility, including Bioconductor,[19] R,[20] Subversion,[21] and GitHub.[22] All of these tools were developed from the bottom up. For example, as researchers were

[19] The Bioconductor website is https://www.bioconductor.org, accessed January 12, 2016.
[20] The R Project website is https://www.r-project.org, accessed January 12, 2016.
[21] The Subversion website is https://subversion.apache.org, accessed January 12, 2016.
[22] The GitHub website is https://github.com, accessed January 12, 2016.

analyzing data, they identified gaps among the available tools and went on to create R and Bioconductor to facilitate their work.

Irizarry stated that it is important to assess if irreproducibility is truly a crisis and if there is a difference now compared to how science was done 50 years ago. For example, the published estimates of the speed of light from 1900 to 1960 were regularly refined, and error bars narrowed (Youden, 1972), which illustrates that the community found a way to continue improving despite the problems that have always surrounded science.

Irizarry was not optimistic about incentive changes such as top-down measures, rules, and regulations, although he agreed with Bourne that data sharing should be incentivized. In addition to enhancing reproducibility, it will encourage more researchers to look at more data and potentially make additional discoveries. However, he cautioned that some policies could be used in unintended ways, and adding hurdles to publication can slow progress in a number of ways.

He noted that although he does not see clear evidence of drastic change in the rate of irreproducibility in the biomedical field, one remarkable change over the past 50 years has been the attention given to press releases, with more emphasis now on getting results in the top newspapers. He also commented that with the quick biomedical transformation from data poor to data driven, much of the infrastructure (people in leadership positions, journal editors, and training programs) has not changed even though the nature of the work has changed dramatically.

To try to help, Irizarry collaborates with as many researchers as possible. He and several of his colleagues have grants funded by NIH's BD2K initiative to create massive open online courses to improve statistics and data analysis among researchers who did not have that as part of their training. Efforts such as these are important in the biomedical sciences and also in other fields that are moving from data poor to data driven.

His final point was that statisticians should not shy away from teaching students how to do applied statistics. This goes beyond teaching methods and theories and includes showing them how to clean and then analyze data, to check and explore the results, and to be skeptical and critical of data analyses. He emphasized that educating researchers who do not have statistics, computing, and reproducibility as part of their formal training is needed to improve that situation.

Jeff Leek, Johns Hopkins University

In his discussion of evidence-based data analysis, Jeff Leek stated that for small to medium-sized problems, reproducibility (if defined as repeatability of analysis) is a solved problem. The tools and ability exist, so it is possible to achieve. Leek noted that the question that remains is why people are not doing it. The main reason he offered is that researchers are not rewarded for it. If senior leadership

in the communities believes this work should be done, it needs to find a way to communicate these ideas and create a suitable environment.

Leek mentioned that many openly available data analysis tools for reproducible research already exist and are being used, such as R Markdown,[23] Galaxy,[24] IPython Notebook,[25] and GitHub.[26] He referenced a recent study he and a collaborator published about the rate at which discoveries are false in science (Jager and Leek, 2014), which ended up stimulating a debate in the scientific literature. Researchers wrote positive and negative responses, reproduced the analysis using the available data and code, and built and improved upon it. However, he noted that reproducibility and replicability work are often unfairly criticized and held to a higher quality standard than original research; this can be a disincentive for researchers interested in conducting this work.

Leek echoed previous speakers in noting that an analysis can be fully reproducible and yet still be wrong (Baggerly and Coombes, 2009). He emphasized that many communities are getting to a point of looking beyond reproducibility to assess if reproducible results are trustworthy. He also agreed that training is often more important than tools yet is often ignored. He commented that he and most of his colleagues are receiving more requests than they can accept to act as statistical referees for papers. Because there are not currently enough trained researchers to fill this role, Leek suggested that the statistics community should think about prevention and ways to (re)train students and researchers quickly. One example of training being scaled up is the Johns Hopkins University Data Science program,[27] which includes a class on reproducible research. This program has trained more than 100,000 people on reproducible research. It includes lessons on data collection and cleaning, exploratory data analysis using GitHub, and version control. This is a program designed specifically for the modern data scientist, who Leek noted is in high demand. It is also important to make clear what kinds of questions researchers are asking, such as whether the data set is analyzed with descriptive, exploratory, or causal inference methods (Leek and Peng, 2015b). Enforcing any statistical procedure, including p-values, across all science would likely result in resentment and mistakes in implementation.

Panel Discussion

Constantine Gatsonis began by asking the panel to comment on the three types of reproducibility, as explained by Victoria Stodden: empirical reproducibil-

[23] The R Markdown website is http://rmarkdown.rstudio.com, accessed January 12, 2016.

[24] The Galaxy website is https://galaxyproject.org, accessed January 12, 2016.

[25] The IPython Notebook website is http://ipython.org/notebook.html, accessed January 12, 2016.

[26] The GitHub website is https://github.com, accessed January 12, 2016.

[27] The Johns Hopkins University Data Science Program website is https://www.coursera.org/specialization/jhudatascience/1, accessed January 12, 2016.

ity, computational reproducibility, and statistical reproducibility. In particular, he suggested that the concepts and challenges of statistical reproducibility—namely, what kinds of evidence are needed to assess reproducibility—have not yet matured as much as they have in discussions of empirical and computational reproducibility. Rafael Irizarry noted that statistical aspects of reproducibility can get very complicated, as was illustrated in some of the case studies and examples discussed throughout the workshop. Much of the understanding of how to best use statistics comes from experience, and often reproducibility is not ensured simply by documenting researchers' methods. He stressed that researchers need to learn from experience how to evaluate and be critical of data analyses. Gatsonis mentioned that at some point, experiences are quantified in the set of assumptions such as what p-value is acceptable. However, there is a debate now over whether the standard p-value of 0.05 is stringent enough to trust the result. Irizarry commented that there is a trade-off between having low false-positive rates and overlooking important discoveries; since both are important, he would prefer that they both increase. A participant countered that lower standards do not increase the rate of discoveries. Instead, the scientific community wants to be sure that the discoveries being reported are actually true. He stated that there are serious costs for false discoveries because people will be following up on misleading results and thereby wasting resources, and increasing standards for publication is not going to slow the rate of true discoveries. Irizarry reiterated his assertion that true positives decrease with the more conservative research standards. Leek commented that much of the discussion around statistical reproducibility involves shifting the p-value up or down or choosing one test statistic over another. He echoed Irizarry's point that the only way to learn how to do good data analysis is to just do it for a while and figure out what works and what does not. He suspects that data analysis needs to be made an empirical discipline whose efficacy can be studied. Bourne agreed that training is an essential component because data science is accelerating what has been going on in computational biology and bioinformatics for some time. A problem he identified is a propagation effect where people without sufficient statistical knowledge apply methodology incorrectly and low-quality analysis proliferates. Education is the only way to curb this.

A participant noted the existence of a generational problem where new data scientists are being trained but there is not a mechanism for current researchers to improve their existing training. Irizarry and Bourne both agreed that ongoing professional development is needed and wanted across fields, and NIH is funding initiatives to support this development. He noted that many of the courses available online are advanced and could be used to fill this need. An example is a program that affords researchers an opportunity to take sabbaticals at highly analytical laboratories to learn techniques that can be applied back in their own laboratories. Leek noted that the NIH-funded course he helped develop is designed for current

researchers. Bourne mentioned that there is also a need for research administrators to gain a better awareness of how things are changing and the importance of this kind of work. Baru added that in his work overseeing NSF's multidirectorate big data program, he has seen that approaches for teaching data management to a geoscientist, for example, are going to be different from those for teaching it to a biologist, a psychologist, or a molecular biologist. He has found that being able to tune curriculum to the audience is important.

A participant commented that the issue of selective inference is the number one problem that hampers statistical replicability. Irizarry agreed and noted that using statistics more appropriately improves results. Bourne also agreed but added that the interdisciplinary nature of scientific research is changing and, while that is not a statistical issue, issues regarding communities and collaboration among them need to be considered.

A participant wondered if there is any information about the backgrounds of the people who are taking the online data science courses such as how many are nonstatistical domain scientists. Much of the material presented in data science master's programs is applicable and important for researchers who would not identify themselves as data scientists. Leek said some data exist through surveys of participants; these surveys indicate a broad community interest in data science, with programs drawing participants from business, economics, and other disciplines. Bourne added that the University of California, San Diego, held a data science workshop that attracted more faculty than any other program at the university. He said that data science is a catalyst to bring together people from diverse disciplines and foster collaborations.

A participant wondered how many years it would take for the community to fully understand reproducibility, especially as it relates to big data. Gatsonis noted that many statistical tools break down in the big data context, and researchers need to think in fresh ways about how to do these types of analyses with large data.

A participant commented that NIH's Gene Expression Omnibus[28] has been a remarkable feat of data sharing: a majority of micro experiments performed have been uploaded, and there is strong buy-in from authors and journals. He wondered if biomedical advances might be slowed due to concerns of privacy when working with sequencing data. Bourne commented that privacy concerns of sharing data are being worked through and discussed. He said that recent policies begin to address some of this, but this issue needs further immediate attention.

A participant noted that the default in large genomic data sets is to resort to multiple hypothesis testing to correct for really small p-values, while keeping the same p-value thresholds, but wondered whether that is a reasonable thing to do.

[28] The NIH's Gene Expression Omnibus website is http://www.ncbi.nlm.nih.gov/geo/, accessed January 12, 2016.

Leek responded that correcting for multiple testing is a good idea, particularly using measures such as the false discovery rate or other error rates, but there are tricky issues when going to higher dimensions in terms of dependence, when to do multiple hypothesis tests, p-value hacking, and selective inference. There are many ways to get things wrong, even if one corrects for multiple testing, but this testing is generally recommended.

References

Allen, L., A. Brand, J. Scott, M. Altman, and M. Hlava. 2014. Publishing: Credit where credit is due. *Nature* 508(7496):312-313.

Alogna, V., M. Attaya, P. Aucoin, Š. Bahník, S. Birch, B. Bornstein, A.R. Birt, et al. 2014. Registered replication report: Schooler & Engstler-Schooler (1990). *Perspectives on Psychological Science* 9(5):556-578.

Altman, M. 2002. A review of JMP 4.03 with special attention to its numerical accuracy. *The American Statistician* 56(1):72-75.

Altman, M. 2008. A fingerprint method for scientific data verification. Pp. 311-316 in *Advances in Computer and Information Sciences and Engineering* (T. Sobh, ed.). Springer, The Netherlands.

Altman, M., and M. Crosas. 2013. The evolution of data citation: From principles to implementation. *IASSIST Quarterly* 37.

Altman, M., and G. King. 2007. A proposed standard for the scholarly citation of quantitative data. *D-lib Magazine* 13(3/4).

Altman, M., and M.P. McDonald. 2001. Choosing reliable statistical software. *Political Science and Politics* 34(03):681-687.

Altman, M., and M.P. McDonald. 2003. Replication with attention to numerical accuracy. *Political Analysis* 11(3):302-307.

Altman, M., L. Andreev, M. Diggory, G. King, E. Kolster, A. Sone, S. Verba, et al. 2001. Overview of the virtual data center project and software. Pp. 203-204 in *JCDL '01: First Joint Conference on Digital Libraries.*

Altman, M., J. Gill, and M.P. McDonald. 2004. Sources of inaccuracy in statistical computation. *Numerical Issues in Statistical Computing for the Social Scientist.* John Wiley & Sons, Hoboken, N.J.

Altman, M., J. Fox, S. Jackman, and A. Zeileis. 2011. A special volume on "Political Methodology." *Journal of Statistical Software* 42(i01).

Azoulay, L., H. Yin, K.B. Filion, J. Assayag, A. Majdan, M.N. Pollak, and S. Suissa. 2012. The use of pioglitazone and the risk of bladder cancer in people with type 2 diabetes: Nested case-control study. *The BJM* 344:e3645.

Baggerly, K., and K.R. Coombes. 2009. Deriving chemosensitivity from cell lines: Forensic bioinformatics and reproducible research in high-throughput biology. *Annals of Applied Statistics* 3(4):1309-1334.

Barni, M., and F. Perez-Gonzalez. 2005. Pushing science into signal processing [my turn]. *IEEE Signal Processing Magazine* 22(4):119-120.

Bayarri, M.J., and A.M. Mayoral. 2002. Bayesian design of 'successful' replications. *The American Statistician* 56:207-214.

Begley, C.G. 2013. Reproducibility: Six red flags for suspect work. *Nature* 497(7450):433-434.

Begley, C.G., and L.M. Ellis. 2012. Drug development: Raise standards for preclinical cancer research. *Nature* 483:531-533.

Begley, C.G., and J.P.A. Ioannidis. 2015. Reproducibility in science: Improving the standard for basic and preclinical research. *Circulation Research* 116(1):116-126.

Begley, S. 2012. In cancer science, many "discoveries" don't hold up. *Reuters*, Health, March 28. http://www.reuters.com/article/2012/03/28/us-science-cancer-idUSBRE82R12P20120328.

Benjamini, Y., and Y. Hochberg. 1995. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B (Methodological)* 57(1):289-300.

Berk, R., L. Brown, A. Buja, K. Zhang, and L. Zhao. 2013. Valid post-selection inference. *Annals of Statistics* 41(2):802-837.

Bernau, C., M. Riester, A.L. Boulesteix, G. Parmigiani, C. Huttenhower, L. Waldron, and L. Trippa. 2014. Cross-study validation for the assessment of prediction algorithms. *Bioinformatics* 30(12):i105-i112.

Berry, D. 2012. Multiplicities in cancer research: Ubiquitous and necessary evils. *Journal of the National Cancer Institute* 104(15):1124-1132.

Berry, D.A. 2007. The difficult and ubiquitous problems of multiplicities. *Pharmaceutical Statistics* 6(3):155-160.

Bezeau, S., and R. Graves. 2001. Statistical power and effect sizes of clinical neuropsychology research. *Journal of Clinical and Experimental Neuropsychology* 23(3):399-406.

Blanken, I., N. van de Ven, M. Zeelenberg, and M.H.C. Meijers. 2014. Three attempts to replicate the moral licensing effect. *Social Psychology* 45(3):223-231.

Boos, D.D., and L.A. Stefanski. 2011. P-value precision and reproducibility. *The American Statistician* 65:213-221.

Bossuyt, P.M., J.B. Reitsma, D.E. Bruns, C.A. Gatsonis, P.P. Glasziou, L.M. Irwig, J.G. Lijmer, D. Moher, D. Rennie, and H.C. de Vet. 2003. Toward complete and accurate reporting of studies of diagnostic accuracy: The STARD initiative. *Academic Radiology* 10(6):664-669.

Brandt, M.J., H. IJzerman, and I. Blanken. 2014. Does recalling moral behavior change the perception of brightness? A replication and meta-analysis of Banerjee, Chatterjee, and Sinha (2012). *Social Psychology* 45(3):246-252.

Buckheit, J., and D.L. Donoho. 1995. Wavelab and reproducible research. Pp. 55-81 in *Wavelets and Statistics* (A. Antoniadis, ed.). Springer, New York, N.Y.

Buja, A., D. Cook. H. Hofmann, M. Lawrence, E.-K. Lee, D.F. Swayne, and H. Wickham. 2009. Statistical inference for exploratory data analysis and model diagnostics. *Philosophical Transactions of the Royal Society A* 367(1906):4361-4383.

Calin-Jageman, R.J., and T.L. Caldwell. 2014. Replication of the Superstition and Performance Study by Damisch, Stoberock, and Mussweiler (2010). *Social Psychology* 45(3):239-245.

Cardwell, C.R., C.C. Abnet, M.M. Cantwell, and L.J. Murray. 2010. Exposure to oral bisphosphonates and risk of esophageal cancer. *Journal of the American Medical Association* 304(6):657-663.

Casella, G., and R.L. Berger. 1987. Reconciling Bayesian and frequentist evidence in the one-sided testing problem. *Journal of the American Statistical Association* 82(397):106-111.

Chan, A.-W., A. Hróbjartsson, M.T. Haahr, P.C. Gøtzsche, and D.G. Altman. 2004. Empirical evidence for selective reporting of outcomes in randomized trials: Comparison of protocols to published articles. *Journal of the American Medical Association* 291(20):2457-2465.

Clayton, J.A., and F.S. Collins. 2014. NIH to balance sex in cell and animal studies. *Nature* 509(7500):282-283.

Cloninger, D.O., and R. Marchesini. 2001. Execution and deterrence: A quasi-controlled group experiment. *Applied Economics* 33:569-576.

Cohen, J. 1962. The statistical power of abnormal-social psychological research: A review. *Journal of Abnormal and Social Psychology* 65:145-153.

Collins, F.S., and L.A. Tabak. 2014. NIH plans to enhance reproducibility. *Nature* 505(7485):612-613.

Colquhoun, D. 2014. An investigation of the false discovery rate and the misinterpretation of p-values. *Royal Society Open Science* 1(3):140216.

Commission on Presidential Debates. 2010. October 17, 2000 Debate Transcript. http://debates.org/index.php?page=october-17-2000-debate-transcript.

Cossins, D. 2014. Setting the record straight. *Scientist* 28(10):48-53.

Couzin-Frankel, J. 2015. Trust me, I'm a medical researcher. *Science* 347(6221):501-503.

Crabbe, J.C., D. Wahlsten, and B.C. Dudek. 1999. Genetics of mouse behavior: Interactions with laboratory environment. *Science* 284(5420):1670-1672.

Dezhbakhsh, H., and J. Shepherd. 2006. The deterrent effect of capital punishment: Evidence from a "judicial experiment." *Economic Inquiry* 44(3):512-535.

Dezhbakhsh, H., P.H. Rubin, and J.M. Shepherd. 2003. Does capital punishment have a deterrent effect? New evidence from post moratorium panel data. *American Law and Economics Review* 5(2):344-376.

Donoho, D.L. 2010. An invitation to reproducible computational research. *Biostatistics* 11(3):385-388.

Donoho, D.L., and X. Huo. 2004. Beamlab and reproducible research. *International Journal of Wavelets, Multiresolution and Information Processing* 2(04):391-414.

Donoho, D.L., A. Maleki, I.U. Rahman, M. Shahram, and V. Stodden. 2009. Reproducible research in computational harmonic analysis. *Computing in Science and Engineering* 11(1):8-18.

Donohue, J.J., and J. Wolfers. 2005. Uses and abuses of empirical evidence in the death penalty debate. *Stanford Law Review* 58(3):791-845.

Donohue, J.J., and J. Wolfers. 2006. The death penalty: No evidence for deterrence. *Economists' Voice.* http://www.deathpenaltyinfo.org/DonohueDeter.pdf.

Doshi, P., T. Jefferson, and C. Del Mar. 2012. The imperative to share clinical study reports: Recommendations from the Tamiflu experience. *PLOS Medicine* 9(4):e1001201.

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

Ehrlich, I. 1975. The deterrent effect of capital punishment: A question of life or death. *American Economic Review* 65(3):397-417.

Errington, T.M., E. Iorns, W. Gunn, F.E. Tan, J. Lomax, and B.A. Nosek. 2014. An open investigation of the reproducibility of cancer biology research. *eLife* 3:e04333.

Esarey, J., A. Wu, R.T. Stevenson, and R.K. Wilson. 2014. Editorial statement. *The Political Methodologist* 22(1):1-2.

Etminan, M., F. Forooghian, J.M. Brophy, S.T. Bird, and D. Maberley. 2012. Oral fluoroquinolones and the risk of retinal detachment. *Journal of the American Medical Association* 307(13):1414-1419.

Fanelli, D. 2010. Do pressures to publish increase scientists' bias? An empirical support from U.S. states data. *PLOS ONE* 5(4):e10271.

Fanelli, D., and J.P.A. Ioannidis. 2013. U.S. Studies may overestimate effect sizes in softer research. *Proceedings of the National Academy of Sciences* 110(37):15031-15036.

Fisher, R.A. 1926. The arrangement of field experiments. *Journal of the Ministry of Agriculture* 33:503-513.

Fisher, R.A. 1935. *The Design of Experiments, II*. Oliver and Boyd, Edinburgh, Scotland.

Freire, J., P. Bonnet, and D. Shasha. 2012. Computational reproducibility: State-of-the-art, challenges, and database research opportunities. Pp. 593-596 in *SIGMOD '12 Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data*.

Funio, E., I. Golani, and Y. Benjamini. 2012. Measuring behavior of animal models: Faults and remedies. *Nature Methods* 9(12):1167-1170.

Furman v. Georgia, 408 US 238 – Supreme Court 1972.

Gardner, T. 2014. A swan in the making. *Science* 345(6199):855.

Garijo, D., S. Kinnings, L. Xie, P.E. Bourne, and Y. Gil. 2013. Quantifying reproducibility in computational biology: The case of the Tuburculosis Drugome. *PLOS ONE* 8(11):e80278.

Garnett, A., M. Altman, L. Andreev, S. Barbarosa, E. Castro, M. Crosas, G. Durand, et al. 2013. Linking OJS and Dataverse. *PKP Scholarly Publishing Conference*. https://pkp.sfu.ca/pkp2013/paper/view/390.

Gelman, A. 2015. "Academics Should Be Made Accountable for Exaggerations in Press Releases about Their Own Work." Blog. Statistical Modeling, Causal Inference, and Social Science. Posted February 22. http://andrewgelman.com/2015/02/22/academics-made-accountable-exaggerations-press-releases-work/.

Gelman, A., and E. Loken. 2014. The statistical crisis in science. *American Scientist* 102(6):460.

Gerber, A.S., and D.P. Green. 2000. The effects of canvassing, telephone calls, and direct mail on voter turnout: A field experiment. *American Political Science Review* 94(03):653-663.

Gibson, C.E., J. Losee, and C. Vitiello. 2014. A replication attempt of stereotype susceptibility (Shih, Pittinsky, and Ambady, 1999): Identity salience and shifts in quantitative performance. *Social Psychology* 45(3):194-198.

Goodfellow, I.J., D. Warde-Farley, M. Mirza, A. Courville, and Y. Bengio. 2013. "Maxout Networks." Pp. 1319-1327 in *Proceedings of the 30th International Conference on Machine Learning*, JMLR Workshop and Conference Proceedings, Volume 28. http://jmlr.csail.mit.edu/proceedings/papers/v28/goodfellow13.pdf.

Goodman, S. 2001. Of *p*-values and Bayes: A modest proposal. *Epidemiology* 12(3):295-297.

Goodman, S.N. 1992. A comment of replication, *p*-values and evidence. *Statistics in Medicine* 11:875-879.

Goodman, S.N., D.G. Altman, and S.L. George. 1998. Statistical reviewing policies of medical journals. *Journal of General Internal Medicine* 13(11):753-756.

Green, J., G. Czanner, G. Reeves, J. Watson, L. Wise, and V. Beral. 2010. Oral bisphosphonates and risk of cancer of esophagus, stomach, and colorectum: Case-control analysis within a UK primary care cohort. *The BMJ* 341:c4444.

Gregg v. Georgia, 428 US 153 – Supreme Court 1976.

Harris, C.R., N. Coburn, D. Rohrer, and H. Pashler. 2013. Two failures to replicate high-performance-goal priming effects. *PLOS ONE* 8(8):e72467.

Hayes, D.N., S. Monti, G. Parmigiani, C.B. Gilks, K. Naoki, A. Bhattacharjee, M.A. Socinski, C. Perou, and M. Meyerson. 2006. Gene expression profiling reveals reproducible human lung adenocarcinoma subtypes in multiple independent patient cohorts. *Journal of Clinical Oncology* 24(31):5079-5090.

Heller, R., and D. Yekutieli. 2014. Replicability analysis for genome-wide association studies. *Annals of Applied Statistics* 8(1):481-498.

Heller, R., M. Bogomolov, and Y. Benjamini. 2014. Deciding whether follow-up studies have replicated findings in a preliminary large-scale omics study. *Proceedings of the National Academy of Sciences* 111(46):16262-16267.

Hill, A.B. 1965. The environment and disease: Association or causation? *Proceedings of the Royal Society of Medicine* 58:295-300.

Hines, W.C., Y. Su, I. Kuhn, K. Polyak, and M.J. Bissell. 2014. Sorting out the FACS: A devil in the details. *Cell Reports* 6(5):779-781.

Hothorn, T., and F. Leisch. 2011. Case studies in reproducibility. *Briefings in Bioinformatics* 12(3):288-300.

Hume, D. 1738. *A Treatise of Human Nature*. Available as a Project Gutenberg EBook. Released February 13, 2010 [EBook #4705], last updated November 10, 2012. https://www.gutenberg.org/files/4705/4705-h/4705-h.htm.

IJzerman, H., I. Blanken, M.J. Brandt, J. M. Oerlemans, M.M.W. Van den Hoogenhof, S.J.M. Franken, and M.W.G. Oerlemans. 2014. Sex differences in distress from infidelity in early adulthood and in later life: A replication and meta-analysis of Shackelford et al. (2004). *Social Psychology* 45(3):202-208.

Imai, K. 2005. Do get-out-the-vote calls reduce turnout? The importance of statistical methods for field experiments. *American Political Science Review* 99(02):283-300.

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

Ioannidis, J.P.A., and M.J. Khoury. 2011. Improving validation practices in "omics" research. *Science* 334(6060):1230-1232.

Ioannidis, J.P.A., R. Tarone, and J.K. McLaughlin. 2011. The false-positive to false-negative ratio in epidemiologic studies. *Epidemiology* 22(4):450-456.

Jager, L.R., and J.T. Leek. 2014. An estimate of the science-wise false discovery rate and application to the top medical literature. *Biostatistics* 15(1):1-12.

John, L.K., G. Lowenstein, and D. Prelec. 2012. Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science* 23(5):524-532.

Johnson, D.J., F. Cheung, and M.B. Donnellan. 2014. Does cleanliness influence moral judgments? A direct replication of Schnall, Benton, and Harvey (2008). *Social Psychology* 45(3):209-215.

Johnson, V.E. 2013a. Uniformly most powerful Bayesian tests. *Annals of Statistics* 41(1):1716-1741.

Johnson, V.E. 2013b. Revised standards for statistical evidence. *Proceedings of the National Academy of Sciences* 110(48):19313-19317.

Karr, A.F. 2014. Why data availability is such a hard problem. *Journal of the International Association for Official Statistics* 30(2):101-107.

Katz, L., S.D. Levitt, and E. Shustorovich. 2003. Prison conditions, capital punishment and deterrence. *American Law and Economics Review* 5(2):318-343.

Kerr, N.L. 1998. HARKing: Hypothesizing after the results are known. *Personality and Social Psychology Review* 2(3):196-217.

King, G. 1995. Replication, replication. *PS: Political Science & Politics* 28(3):444-452.

Kiselycznyk, C., and A. Holmes. 2011. All (C57BL/6) mice are not created equal. *Frontiers in Neuroscience* 5(10).

Klein, R.A., K.A. Ratliff, M.Vianello, R.B. Adams, Jr., Š. Bahník, M.J. Bernstein, K. Bocian, et al. 2014. Investigating variation in replicability: A "many labs" replication project. *Social Psychology* 45(3):142-152.

Laine, C., S.N. Goodman, M.E. Griswold, and H.C. Sox. 2007. Reproducible research: Moving toward research the public can really trust. *Annals of Internal Medicine* 146(6):450-453.

Lambert, D., and W.J. Hall. 1982. Asymptotic lognormality of p-values. *Annals of Statistics* 10:44-64.

Lash, T.L. 2015. Truth and consequences. *Epidemiology* 26(2):141-142.

Leamer, E.E. 1978. *Specification Searches: Ad Hoc Inference with Nonexperimental Data*. Wiley & Sons, New York, N.Y.

Lee, C.-Y., S. Xie, P.W. Gallagher, Z. Zhang, and Z. Tu. 2015. Deeply supervised nets. *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, Vol. 38. http://jmlr.org/proceedings/papers/v38/lee15a.pdf.

Leek, J.T., and R.D. Peng. 2015a. What is the question? *Science* 347(6228):1314-1315.

Leek, J.T., and R.D. Peng. 2015b. Opinion: Reproducible research can still be wrong: Adopting a prevention approach. *Proceedings of the National Academy of Sciences* 112(6):1645-1646.

Lehrer, J. 2010. The truth wears off: Is there something wrong with the scientific method? *The New Yorker*, December 13.

LeVeque, R.J. 2013. "Top ten reasons to *not* share your code (and why you should anyway). *SIAM News*. http://sinews.siam.org/DetailsPage/tabid/607/ArticleID/386/Top-Ten-Reasons-To-Not-Share-Your-Code-and-why-you-should-anyway.aspx.

LeVeque, R.J., I.M. Mitchell, and V. Stodden. 2012. Reproducible research for scientific computing: Tools and strategies for changing the culture. *Computing in Science and Engineering* 14(4):13.

Li, Q., J.B. Brown, H. Huang, and P.J. Bickel. 2011. Measuring reproducibility of high-throughput experiments. *Annals of Applied Statistics* 5(3):1752-1779.

Liberati, A., D.G. Altman, J. Tetzlaff, C. Mulrow, P.C. Gøtzsche, J.P.A. Ioannidis, M. Clarke, P.J. Devereaux, J. Kleijnen, and D. Moher. 2009. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: Explanation and elaboration. *Annals of Internal Medicine* 151(4):W65-W94.

Lynott, D., K.S. Corker, J. Wortman, L. Connell, M.B. Donnellan, R.E. Lucas, and K. O'Brien. 2014. Replication of "Experiencing physical warmth promotes interpersonal warmth" by Williams and Bargh (2008). *Social Psychology* 45(3):216-222.

Madigan, D., P.E. Stang, J.A. Berlin, M. Schuemie, J.M. Overhang, M.A. Suchard, B. Dumouchel, A.G. Hartzema, and P.B. Ryan. 2014. A systemic statistical approach to evaluating evidence from observational studies. *Annual Review of Statistics and Its Applications* 1:11-39.

Makel, M.C., J.A. Plucker, and B. Hegarty. 2012. Replications in psychology research; how often do they really occur? *Perspectives on Psychological Science* 7(6):537-542.

Mann, C.C. 1994. Behavioral genetics in transition. *Science* 265(5166):1686-1689.

McNutt, M. 2014. Journals unite for reproducibility. *Science* 346(6210):679.

Miller, R.G. 1986. *Beyond ANOVA, Basics of Applied Statistics*. Wiley, New York.

Mills, J.L. 1993. Data torturing. *New England Journal of Medicine* 329:1196-1199.

Mobley, A., S.K. Linder, R. Braeuer, L.M. Ellis, and L. Zwelling. 2013. A survey on data reproducibility in cancer research provides insights into our limited ability to translate findings from the laboratory to the clinic. *PLOS ONE* 8(5):e63221.

Mocan, H.N., and R.K. Gittings. 2003. Getting off death row: Commuted sentences and the deterrent effect of capital punishment. *Journal of Law and Economics* XLVI:453-478.

Mojirsheibani, M., and R. Tibshirani. 1996. Some results on bootstrap prediction intervals. *Canadian Journal of Statistics* 24:549-568.

Molina, H., G. Parmigiani, and A. Pandey. 2005. Assessing reproducibility of a protein dynamics study using in vivo labeling and liquid chromatography tandem mass spectrometry. *Analytical Chemistry* 77(9):2739-2744.

Moon, A., and S.S. Roeder. 2014. A secondary replication attempt of stereotype susceptibility (Shih, Pittinsky, and Ambady, 1999). *Social Psychology* 45(3):199-201.

Moore, J. 2013. "Onstage Speech Transcript: Actress in a Leading Role." Transcript. http://www.oscars.org/press/onstage-speech-transcript-actress-leading-role.

Mosteller, F., and J.W. Tukey. 1977. *Data Analysis and Regression: A Second Course in Statistics*. Addison-Wesley, Reading, Mass.

Motulsky, H.J. 2014. Common misconceptions about data analysis and statistics. *British Journal of Pharmacology* 172(8):2126-2132.

Müller, F., and K. Rothermund. 2014. What does it take to activate stereotypes? Simple primes don't seem to be enough: A replication of stereotype activation (Banaji and Hardin, 1996; Blair and Banaji, 1996). *Social Psychology* 45(3):187-193.

NASEM (National Academies of Sciences, Engineering, and Medicine). 2015. *Reproducibility Issues in Research with Animals and Animal Models*. The National Academies Press, Washington, D.C.

Nature Methods. 2012. "All Things Being Equal." Editorial. 9(2):111.

Nature Neuroscience. 2013. "Raising Standards." Editorial. 16(5):517.

Nauts, S., O. Langner, I. Huijsmans, R. Vonk, and D.H.J. Wigboldus. 2014. Forming impressions of personality: A replication and review of Asch's (1946) Evidence for a primacy-of-warmth effect in impression formation. *Social Psychology* 45(3):153-163.

Netzer, Y., T. Wang, A. Coates, A. Bissacco, B. Wu, and A.Y. Ng. 2011. "Reading Digits in Natural Images with Unsupervised Feature Learning." Deep Learning and Un-supervised Feature Learning Workshop, NIPS. http://ufldl.stanford.edu/housenumbers/nips2011_housenumbers.pdf.

Nicholson, J., and Y. Lazebnik. 2014. "The R-Factor: A Measure of Scientific Veracity." *The Winnower*. https://thewinnower.com/papers/1-the-r-factor-a-measure-of-scientific-veracity.

Nosek, B.A. 2014. Registered reports: A method to increase the credibility of published results. *Social Psychology* 45(3):137-141.

Nosek, B.A., and D. Lakens. 2014. Registered reports. *Social Psychology* 45(3):137-141.

Nosek, B.A., J.R. Spies, and M. Motyl. 2012. Scientific utopia II. Restructuring incentives and practices to promote truth over publishability. *Perspectives on Psychological Science* 7(6):615-631.

Nosek, B.A., G. Alter, G.C. Banks, D. Borsboom, S.D. Bowman, S.J. Breckler, S. Buck, et al. 2015. Promoting an open research culture. *Science* 348(6242):1422-1425.

NRC (National Research Council). 1966. *Languages and Machines: Computers in Translation and Linguistics*. National Academy Press, Washington, D.C.

NRC. 1978. *Deterrence and Incapacitation: Estimating the Effects of Criminal Sanctions on Crime Rates*. National Academy Press, Washington, D.C.

NRC. 1991. *The Future of Statistical Software: Proceedings of a Forum*. National Academy Press, Washington, D.C.

NRC. 2012. *Deterrence and the Death Penalty*. The National Academies Press, Washington, D.C.

NSF (National Science Foundation). 2014. "A Framework for Ongoing and Future National Science Foundation Activities to Improve Reproducibility, Replicability, and Robustness in Funded Research." Prepared for the Office of Management and Budget. Submitted December 31, 2014. https://www.nsf.gov/attachments/134722/public/Reproducibility_NSFPlanforOMB_Dec31_2014.pdf.

NSF. 2015. *Social, Behavioral, and Economic Sciences Perspectives on Robust and Reliable Science*. Report of the Subcommittee on Replicability in Science Advisory Committee to the National Science Foundation Directorate for Social, Behavioral, and Economic Sciences. May. http://www.nsf.gov/sbe/AC_Materials/SBE_Robust_and_Reliable_Research_Report.pdf.

Obama, B. 2006. *The Audacity of Hope: Thoughts on Reclaiming the American Dream*. Random House Crown Publishing, New York, N.Y.

Open Science Collaboration. 2015. Estimating the reproducibility of psychological science. *Science* 349(6251):943-951.

Pallet, D.S. 1985. Performance assessment of automatic speech recognizers. *Journal of Research of the National Bureau of Standards* 90(5):371-387.

Pashler, H., N. Coburn, and C.R. Harris. 2012. Priming of social distance? Failure to replicate effects on social and food judgments. *PLOS ONE* 7(8):e42510.

Pasternak, B., H. Svanström, M. Melbye, and A. Hviid. 2013. Association between oral fluoroquinolone use and retinal detachment. *Journal of the American Medical Association* 310(20):2184-2190.

Patel, C.J., and J.P.A. Ioannidis. 2014. Studying the elusive environment in large scale. *Journal of the American Medical Association* 311(21):2173-2174.

Patel, C.J., B. Burford, and J.P.A. Ioannidis. 2015. Assessment of vibration of effects due to model specification can demonstrate the instability of observational associations. *Journal of Clinical Epidemiology* 68 (2015):1046-1058.

Peers, I.S., P.R. Ceuppens, and C. Harbron. 2012. In search of preclinical robustness. *Nature Reviews Drug Discovery* 11(10):733-734.

Peng, R. 2009. Reproducible research and biostatistics. *Biostatistics* 10(3):405-408.

Peng, R. 2011. Reproducible research in computational science. *Science* 334(6060):1226-1227.

Peng, R.D., F. Dominici, and S.L. Zeger. 2006. Reproducible epidemiology research. *American Journal of Epidemiology* 163(9):783-789.

Pereira, T.V., R.I. Horwitz, and J.P.A. Ioannidis. 2012. Empirical evaluation of very large treatment effects of medical intervention. *Journal of the American Medical Association* 308(16):1676-1684.

Perrin, S. 2014. Preclinical research: Make mouse studies work. *Nature* 507(7493):423-425.

Pierce, J.R. 1969. Whither speech recognition. *Journal of the Acoustical Society of America* 46:1049-1050.

Piwowar, H., R.S. Day, and D.B. Fridsma. 2007. Sharing detailed research data is associated with increased citation rate. *PLOS ONE* 2(3):e308.

Political Science Journal Editors. 2014. "Data Access and Research Transparency (DA-RT): A Joint Statement." http://www.dartstatement.org/.

Prinz, F., T. Schlange, and K. Asadullah. 2011. Believe it or not: How much can we rely on published data on potential drug targets? *Nature Reviews Drug Discovery* 10:712.

Rasko, J., and C. Power. 2015. What pushes scientists to lie? The disturbing but familiar story of Haruko Obokata. *The Guardian,* February 18.

Redelmeier, D.A., and S.M. Singh. 2001. Survival in Academy Award-winning actors and actresses. *Annals of Internal Medicine* 134:955-962.

Reiter, J.P., and S.K. Kinney. 2011. Sharing confidential data for research purposes: A primer. *Epidemiology* 22(5):632-635.

Rekdal, O.B. 2014. Academic urban legends. *Social Studies of Science* 44(4):638-654.

Richter, S.H., J.P. Garner, and H. Würbel. 2009. Environmental standardization: Cure or cause of poor reproducibility in animal experiments. *Nature Methods* 6:257-261.

Richter, S.H., J.P. Garner, B. Zipser, L. Lewejohann, N. Sachser, C. Touma, B. Schindler, et al. 2011. Effect of heterogenization on the reproducibility of mouse behaviour: A multi-laboratory study. *PLOS ONE* 6(1):e16461.

Rosenthal, R. 1979. The file drawer problem and tolerance for null results. *Psychological Bulletin* 86(3):638-641.

Rothman, K.J., and J.D. Boice, Jr. 1979. *Epidemiologic Analysis with a Programmable Calculator.* NIH Publication 79-1649. U.S. Government Printing Office, Washington, D.C.

Ryan, P.B., P.E. Stang, J.M. Overhage, M.A. Succhard, A.G. Hartzema, W. DuMouchel, C.G. Reich, M.J. Schuemie, and D. Madigan. 2013. A comparison of the empirical performance of methods for a risk identification system. *Drug Safety* 36(1):143-158.

Schooler, J.W. 2014. Turning the lens of science on itself: Verbal overshadowing, replication, and metascience. *Perspectives on Psychological Science* 9(5):579-584.

Schuemie, M.J., P.B. Ryan, W. DuMouchel, M.A. Suchard, and D. Madigan. 2013. Interpreting observational studies; why empirical calibration is needed to correct p-values. *Statistics in Medicine* 33(2):209-218.

Scott, S., J.E. Kranz, J. Cole, J.M. Lincecum, K. Thompson, N. Kelly, A. Bostrom, et al. 2008. Design, power, and interpretation of studies in the standard murine model of ALS. *Amyotrophic Lateral Sclerosis* 9(1):4-15.

Sedlmeier, P., and G. Gigerenzer. 1989. Do studies of statistical power have an effect on the power of studies? *Psychological Bulletin* 105:309-316.

Sellin, T. 1959. *The Death Penalty*. American Law Institute, Philadelphia, Pa.

Sena, E., H.B. van der Worp, D. Howells, and M. Macleod. 2007. How can we improve the pre-clinical development of drugs for stroke? *Trends in Neuroscience* 30:433-439.

Shao, J., and S.-C. Chow. 2002. Reproducibility probability in clinical trials. *Statistics in Medicine* 21:1727-1742.

Shao, J., and D. Tu. 1996. *The Jackknife and Bootstrap.* Springer, New York, N.Y.

Shepherd, J.M. 2005. Deterrence versus brutalization: Capital punishment's differing impacts among states. *Michigan Law Review* 104:248.

Simmons, J.P., N.D. Nelson, and U. Simonsohn. 2011. False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science* 22(11):1359-1366.

Simons, D.J., A.O. Holcombe, and B.A. Spellman. 2014. An introduction to registered replication reports at perspectives on psychological science. *Perspectives on Psychological Science* 9(5):552-555.

Simonsohn, U. 2012. It does not follow: Evaluating the one-off publication bias critiques by Francis (2012a, 2012b, 2012c, 2012d, 2012e, in press). *Perspectives on Psychological Science* 7(6):597-599.

Sinclair, H.C., K.B. Hood, and B.L. Wright. 2014. Revisiting the Romeo and Juliet effect (Driscoll, Davis, and Lipetz, 1972): Reexamining the links between social network opinions and romantic relationship output. *Social Psychology* 45(3):170-178.

Spence, D. 2014. Evidence based medicine is broken. *The BMJ* 348.

Steward, O., P.G. Popovich, W.D. Dietric, and N. Kleitman. 2012. Replication and reproducibility in spinal cord injury. *Experimental Neurology* 233(2):597-605.

Stodden, V. 2009a. Enabling reproducible research: Open licensing for scientific innovation. *International Journal of Communications Law and Policy*. March 3. Available at Social Science Electronic Publishing, http://ssrn.com/abstract=1362040.

Stodden, V. 2009b. The legal framework for reproducible scientific research: Licensing and copyright. *Computing in Science and Engineering* 11(1):35-40.

Stodden, V. 2013. Resolving irreproducibility in empirical and computational research. *IMS Bulletin Online*. http://bulletin.imstat.org/2013/11/resolving-irreproducibility-in-empirical-and-computational-research/.

Stodden, V., J. Borwein, and D. Bailey. 2013a. "Setting the default to reproducible" in computational science research. *SIAM News* 46:4-6.

Stodden, V., P. Guo, and Z. Ma. 2013b. Toward reproducible computational research: An empirical analysis of data and code policy adoption by journals. *PLOS ONE* 8(6):e67111.

Stodden, V., F. Leisch, and R.D. Peng. 2014. *Implementing Reproducible Research*. CRC Press, Boca Raton, Fla.

Stodden, V., S. Miguez, and J. Seiler. 2015. ResearchCompendia.org: Cyberinfrastructure for reproducibility and collaboration in computational science. *Computing in Science and Engineering* 17(1):12-19.

Sylvestre, M.-P., E. Husztl, and J.A. Hanley. 2006. Do Oscar winners live longer than less successful peers? A reanalysis of the evidence. *Annals of Internal Medicine* 145:361-363.

Terrorist Penalties Enhancement Act of 2003: Hearing on H.R. 2934 Before the Subcomm. on Crime, Terrorism, and Homeland Security of the H. Comm. on the Judiciary, 108th Cong. 10-11 (2004). http://commdocs.house.gov/committees/judiciary/hju93224.000/hju93224_0f.htm.

Trafimow, D., and M. Marks. 2015. Editorial. *Basic and Applied Social Psychology* 37:1-2.

Vandewalle, P., J. Kovacevic, and M. Vetterli. 2009. Reproducible research in signal processing—What, why, and how. *IEEE Signal Processing Magazine* 26(3):37-47.

Vermeulen, I., A. Batenburg, C.J. Beukeboom, and T. Smits. 2014. Breakthrough or one-hit wonder? Three attempts to replicate single-exposure musical conditioning effects on choice behavior (Gorn, 1982). *Social Psychology* 45(3):179-186.

Wahlsten, D. 2001. Standardized tests of mouse behavior: Reasons, recommendations, and reality. *Physiology Behavior* 73(5):695-704.

Waldron, L., B. Haibe-Kains, A.C. Culhane, M. Riester, J. Ding, X.V. Wang, M. Ahmadifar, et al. 2014. Comparative meta-analysis of prognostic gene signatures for late-stage ovarian cancer. *Journal of the National Cancer Institute* 106(5):dju049.

Wei, L., T.M. MacDonald, and I.S. Mackenzie. 2013. Pioglitazone and bladder cancer: A propensity score matched cohort study. *Journal of Clinical Pharmacology* 75(1):254-259.

Wellcome Trust. 2014. *Establishing Incentives and Changing Cultures to Support Data Access*. May. http://www.wellcome.ac.uk/stellent/groups/corporatesite/@msh_peda/documents/web_document/wtp056495.pdf.

Wesselmann, E.D., K.D. Williams, J.B. Pryor, F.A. Eichler, D.M. Gill, and J.D. Hogue. 2014. Revisiting Schachter's research on rejection, deviance, and communication (1951). *Social Psychology* 45(3):164-169.

White, H. 2000. A reality check for data snooping. *Econometrica* 68(5):1097-1126.

Wicherts, J.M., D. Borsboom, J. Kats, and D. Molenaar. 2006. The poor availability of psychological research data for reanalysis. *American Psychologist* 61(7):726-728.

Wilson, E.O. 1998. *Consilience: The Unity of Knowledge*. Random House, New York, N.Y.

Wolfinger, R.D. 2013. Reanalysis of Richter et al. (2010) on reproducibility. *Nature Methods* 10:373-374.

Würbel, H., S.H. Richter, and J.P. Garner. 2013. Reply to: "Reanalysis of Richter et al. (2010) on reproducibility." *Nature Methods* 10:374.

Yale Roundtable Participants. 2010. Reproducible research. *Computing in Science and Engineering* 12(5):8-13.

Youden, W.J. 1972. Enduring values. *Technometrics* 14(1):1-14.

Young, S.S., and A. Karr. 2011. Deming, data, and observational studies: A process out of control and needing fixing. *Significance* 8(3):116-120.

Žeželj, I.L., and B.R. Jokić. 2014. Replication of experiments evaluating impact of psychological distance on moral judgment (Eyal, Liberman and Trope, 2008; Gong and Medin, 2012). *Social Psychology* 45(3):223-231.

Zilliox, M.J., and R. Irizarry. 2007. A gene expression barcode for microarray data. *Nature Methods* 4(11):911-913.

Zimmerman, P.R. 2004. State executions, deterrence, and the incidence of murder. *Journal of Applied Economics* VII(I):163-193.

Appendixes

A

Registered Workshop Participants

Altman, Micah – Massachusetts Institute of Technology
Amarreh, Ishmael – National Institutes of Health
Andrews, Howard – Columbia University Mailman School of Public Health
Anestidou, Lida – National Academies of Sciences, Engineering, and Medicine
Arora, Vipin – U.S. Energy Information Administration
Arrieta, Dan – Chevron Phillips Chemical Company LP
Arrison, Tom – National Academies of Sciences, Engineering, and Medicine
Auh, Sungyoung – National Institutes of Health
Baggerly, Keith – MD Anderson Cancer Center
Barberia, Lorena – University of São Paulo, Brazil
Baru, Chaitan – National Science Foundation
Bastian, Hilda – National Institutes of Health
Beaton, Maura – National Health Service
Beck, Nancy – American Chemistry Council
Becker, Chandler – National Institute of Standards and Technology
Benjamini, Yoav – Tel Aviv University, Israel
Berman, Lew – ICF International
Bhat, Talapady – National Institute of Standards and Technology
Boehm, Frederick – University of Wisconsin, Madison
Bogoslovsky, Tanya – Uniformed Services University of the Health Sciences
Boisvert, Ronald – National Institute of Standards and Technology
Boos, Dennis – North Carolina State University
Bourne, Philip – National Institutes of Health

Brooks, Frank – Mallinckrodt Institute of Radiology
Buck, Stuart – Laura and John Arnold Foundation
Buja, Andreas – Wharton School of the University of Pennsylvania
Campbell, Greg – U.S. Food and Drug Administration
Carney, Joanne – American Association for the Advancement of Science
Casano, Patricia Kablach – General Electric Company
Ceneviva, Ricardo – Rio de Janeiro State University, Brazil
Chen, Jinbo – University of Pennsylvania
Chen, Nanyu – LinkedIn Corporation
Chen, Quan – National Institutes of Health
Chiuzan, Cody – Columbia University
Christensen, Garret – Berkeley Initiative for Transparency in the Social Sciences
Clark, Jennifer – U.S. Food and Drug Administration
Cockerill, Matthew – Riffyn, Inc.
Cohen, Michael – National Academies of Sciences, Engineering, and Medicine
Cohen, Michael P. – American Institutes for Research
Costello, Donald – University of Nebraska
Cragin, Melissa – National Science Foundation
Creel, Darryl – RTI International
Davidson, Sara – Korea Transport Institute
Dobbins, Janet – The Institute for Statistics Education at Statistics.com
Doll, Nancy – Precise Clinical Data Collection
Duan, Fenghai – Brown University
Duan, Weitao – LinkedIn Corporation
DuBreuil, Dan – Brown University
Dunn, Michelle – National Institutes of Health
Eisenberg, Jon – National Academies of Sciences, Engineering, and Medicine
Elliott, John – National Institute of Standards and Technology
Errington, Tim – Center for Open Science
Esarey, Justin – Rice University
Facelli, Julio – University of Utah
Fienberg, Steve – Carnegie Mellon University
Fitzmaurice, J. Michael – Agency for Healthcare Research and Quality
Florance, Valerie – National Library of Medicine
Forde, Kimberly – University of Pennsylvania
French, Benjamin – University of Pennsylvania
Fritts, Martin – Leidos Biomedical Research
Gage, Joseph – University of Wisconsin, Madison
Gardner, Timothy – Riffyn, Inc.
Gatsonis, Constantine – Brown University
Ge, Ping – U.S. Department of Energy

Goldberg, Judith – New York University School of Medicine
Goodman, Steve – Stanford University
Gordon, Ruthanna – AVIAN Engineering
Guthrie, William – National Institute of Standards and Technology
Hardin, Johanna – Pomona College
Harnly, James – U.S. Department of Agriculture
Harris, Richard – National Public Radio
Ho, Jeff – Stanford University
Holmes, John – University of Pennsylvania Perelman School of Medicine
Huo, Xiaoming – National Science Foundation
Inupakutika, Devasena – University of Southampton
Ioannidis, John – Stanford University
Irizarry, Rafael – Harvard University
Jalalpour, Mehdi – Cleveland State University
Jasny, Barbara – Science/AAAS
Jeanotte, Alexis – AVIAN, LLC
Jia, Haomiao – Columbia University
Johnson, Valen – Texas A&M University
Kapatou, Alexandra – American University
Karr, Alan – RTI International
Kass, Robert – Carnegie Mellon University
Kaufhold, John – Deep Learning Analytics
Kelly, Anthony – National Science Foundation
Khadka, Alla – University of Pittsburgh
Korolev, Vladimir – University of Maryland, Baltimore County
Kovacevic, Jelena – Carnegie Mellon University
Kozlovski, Tal – Tel Aviv University, Israel
LaLonde, Donna – American Statistical Association
Lazebnik, Yuri – Cancer Biology Consultant
Lee, Meredith – Homeland Security Advanced Research Projects Agency
Leek, Jeff – Johns Hopkins University
LeVeque, Randy – University of Washington
Levermore, David – University of Maryland
Li, Jingyi Jessica – University of California, Los Angeles
Li, Qianqiu – Jansen Research and Development
Li, Qunhua – Pennsylvania State University
Li, Shanshan – National Institutes of Health
Liao, K. George – Wildfire Analytics and Research
Liberman, Mark – University of Pennsylvania
Lin, Xihong – Harvard University
Locascio, Laurie – National Institute of Standards and Technology

Lomax, Joelle – Science Exchange
Lu, John – National Institute of Standards and Technology
Lucas, Philip – National Institutes of Health
Lystig, Theodore – Medtronic
Macleod, Malcolm – University of Edinburgh, Scotland
Manrai, Arjun – Harvard Medical School
Marcus, Stephen – National Institutes of Health
Markatou, Marianthi – State University of New York at Buffalo
McGuire, Susan – National Institutes of Health
McLaughlin, Gerald – National Institutes of Health
McNutt, Marcia – *Science* magazine
Meacham, Steve – National Science Foundation
Meiring, Wendy – University of California, Santa Barbara
Messinger-Cayetano, Shari – University of Miami
Mitelpunkt, Alexis – Tel Aviv University, Israel
Mouakkad, Sally – Research Councils UK
Mukhopadhyay, Rajendrani – American Society for Biochemistry and Molecular Biology
Nelson, Stuart – Retired
Novikova, Irina – World Bank
Nuzzo, Regina – Gallaudet University
Olson, Nate – National Institute of Standards and Technology
Olster, Deborah – National Science Foundation
Ord, Keith – Georgetown University
Pai, Vinay – National Institutes of Health
Panagiotou, Orestis – National Institutes of Health
Pantula, Sastry – Oregon State University
Parikh, Hemang – National Institute of Standards and Technology
Park, Young – San Jose State University
Parmigiani, Giovanni – Dana-Farber Cancer Institute
Pavlick, Andrea – AAAS/U.S. Environmental Protection Agency
Pearl, Jennifer – National Science Foundation
Peng, Roger – Johns Hopkins Bloomberg School of Public Health
Percival, Matthew
Perfito, Nicole – Science Exchange
Pierson, Steve – American Statistical Association
Qualters, Irene – National Science Foundation
Radman, Thomas – National Institutes of Health
Rajapakse, Sula – TFI
Ramstein, Guillaume – University of Wisconsin, Madison
Raphael, Louise – Howard University

Ratpan, Flora – NOVA Chemicals
Ray, Rebecca – University of Wisconsin, Madison
Risso, Davide – University of California, Berkeley
Robinowitz, Max – Retired
Rosemond, Erica – National Institutes of Health
Russek-Cohen, Estelle – U.S. Food and Drug Administration
Russell, Adam – Intelligence Advanced Research Projects Activity
Russell, Thomas – National Science Foundation
Ryan, Patrick – Janssen Research and Development
Sadsad, Rosemarie – Westmead Hospital
Saigal, Sanjay – University of California, Davis
Salnikow, Konstantin – National Institutes of Health
Schneeman, Paul – American Statistical Association
Schuette, Paul – U.S. Food and Drug Administration
Schwalbe, Michelle – National Academies of Sciences, Engineering, and Medicine
Scovell, James – Intel
Sebastian, Rhonda – U.S. Department of Agriculture
Selimovic, Seila – U.S. Department of State
Setti, Gianluca – University of Ferrara, Italy
Sezgin, Efe – Johns Hopkins University
Shachar, Netta – Tel Aviv University, Israel
Shieh, Ching-yi – National Institutes of Health
Soderberg, Courtney – Center for Open Science
Sorace, James – U.S. Department of Health and Human Services
Sorkin, Barbara – National Institutes of Health
Stodden, Victoria – University of Illinois, Urbana-Champaign
Straf, Miron – Virginia Bioinformatics Institute, Virginia Tech
Subramanian, Ayshwarya – Harvard School of Public Health
Suchard, Marc – University of California, Los Angeles
Sun, Hong-Wei – National Institutes of Health
Szewczyk, William – National Security Agency
Tabak, Lawrence – National Institutes of Health
Tannouri, Ahlam – Morgan State University
Thompson, Paul – Sanford Research
Tong, Frances – BD Technologies
Triche, Tim – University of Southern California
Tuttle, Mark
Vaagenes, Ian – Hines VA Hospital
Valeri, Linda – Harvard School of Public Health
Vargas, Juan
Vattikuti, Shashaank – National Institutes of Health

Waldron, Levi – City University of New York School of Public Health
Wang, Eva – HTG Molecular Diagnostics
Warshavski, Dan – Tel Aviv University, Israel
Wasserstein, Ronald – American Statistical Association
Waugh, Shawna – U.S. Energy Information Administration
Whittington, Kjersten – National Institutes of Health
Wolfers, Justin – University of Michigan
Wood, Benjamin – International Initiative for Impact Evaluation
Wright, Tracy – U.S. Environmental Protection Agency
Xia, Ashley – National Institutes of Health
Xiao, Ying – Thomas Jefferson University
Xu, Han – Dana-Farber Cancer Institute
Xu, Ya – LinkedIn Corporation
Yaseen, Muhammad – University of Agriculture Faisalabad, Pakistan
Yuan, May – University of Texas, Dallas
Zarin, Deborah – National Library of Medicine
Zhao, Fen – National Science Foundation
Zhu, Yiliang – University of South Florida

B

Workshop Agenda

DAY 1
FEBRUARY 26, 2015

Session I: Overview and Case Studies

8:30 a.m. **Introductions from the Workshop Co-Chairs**

Constantine Gatsonis, Brown University
Giovanni Parmigiani, Dana-Farber Cancer Institute

8:45 **Perspectives from Stakeholders**

Lawrence Tabak, National Institutes of Health
Irene Qualters, National Science Foundation
Justin Esarey, Rice University and *The Political Methodologist*
Gianluca Setti, University of Ferrara, Italy, and IEEE
Joelle Lomax, Science Exchange

9:45 **Overview of the Workshop**

Victoria Stodden, University of Illinois, Urbana-Champaign

10:30 **Case Studies**

Yoav Benjamini, Tel Aviv University
Justin Wolfers, University of Michigan

Session II: Conceptualizing, Measuring, and Studying Reproducibility

1:30 p.m. **Definitions and Measures of Reproducibility**

Speaker: Steven Goodman, Stanford University
Discussant: Yoav Benjamini, Tel Aviv University

2:30 **Reproducibility and "Statistical Significance"**

Speaker: Dennis Boos, North Carolina State University
Discussants: Andreas Buja, Wharton, University of Pennsylvania
Val Johnson, Texas A&M University

3:45 **Assessment of Factors Affecting Reproducibility**

Speaker: Marc Suchard, University of California, Los Angeles
Discussants: Courtney Soderberg, Center for Open Science
John Ioannidis, Stanford University

4:45 **Reproducibility from the Informatics Perspective**

Speaker: Mark Liberman, University of Pennsylvania
Discussant: Micah Altman, Massachusetts Institute of Technology

5:45 **Adjourn Day 1**

DAY 2
FEBRUARY 27, 2015

Session III: The Way Forward: Using Statistics to Achieve Reproducibility

8:30 a.m. **Panel Discussion: Open Problems, Needs, and Opportunities for Methodologic Research**

Moderator: Giovanni Parmigiani, Dana-Farber Cancer Institute
Panelists: Lida Anestidou, National Academies of Sciences, Engineering, and Medicine
Tim Errington, Center for Open Science
Xiaoming Huo, National Science Foundation
Roger Peng, Johns Hopkins Bloomberg School of Public Health

10:00 **Panel Discussion: Reporting Scientific Results and Sharing Scientific Study Data**

Moderator: Victoria Stodden, University of Illinois, Urbana-Champaign
Panelists: Keith Baggerly, MD Anderson Cancer Center
Ronald Boisvert, Association for Computing Machinery and National Institute of Standards and Technology
Randy LeVeque, Society for Industrial and Applied Mathematics and University of Washington
Marcia McNutt, *Science* magazine

11:45 **Panel Discussion: The Way Forward from the Data Sciences Perspective: Research**

Moderator: Constantine Gatsonis, Brown University
Panelists: Chaitan Baru, National Science Foundation
Philip Bourne, National Institutes of Health
Rafael Irizarry, Harvard University
Jeff Leek, Johns Hopkins University

1:00 p.m. **Adjourn Workshop**

C

Acronyms

ACI	Advanced Cyberinfrastructure
ACM	Association for Computing Machinery
BD2K	Big Data to Knowledge (project)
CATS	Committee on Applied and Theoretical Statistics
DARPA	Defense Advanced Research Projects Agency
DMS	Division of Mathematical Sciences
DoD	Department of Defense
DOI	digital object identifier
GxL	genotype *x* laboratory
IACUC	Institutional Animal Care and Use Committee
ICERM	Institute for Computational and Experimental Research in Mathematics
IEEE	Institute of Electrical and Electronics Engineers
ILAR	Institute for Laboratory Animal Research
IRB	Internal Review Board
NASEM	National Academies of Sciences, Engineering, and Medicine
NIH	National Institutes of Health
NIST	National Institute of Standards and Technology

NRC	National Research Council
NSF	National Science Foundation
OMOP	Observational Medical Outcomes Partnership
SIAM	Society for Industrial and Applied Mathematics
SIGMOD	Special Interest Group on the Management of Data
SPS	Signal Processing Society
TOMS	Transactions on Mathematical Software
TSCS	time series cross section
UMPBT	uniformly most powerful Bayesian test
XSEDE	Extreme Science and Engineering Discovery Environment