REPRODUCIBILITY, REPLICABILITY, AND GENERALIZATION IN THE SOCIAL, BEHAVIORAL, AND ECONOMIC SCIENCES

Report of the Subcommittee on Replicability in Science of the SBE Advisory Committee to the National Science Foundation
13 May 2015 Presentation
at SBE AC Spring Meeting by K. Bollen.

Report of the Subcommittee on Replicability in Science

Advisory Committee to the NSF SBE Directorate

Subcommittee Members:

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

Staff Assistance

Heather Dean (National Science Foundation)

TOPICS

I. Background
II. Subcommittee Report
III. Definitions
IV. Recommendations
V. Final Comments

Background

- Spring, 2013
  - NSF SBE Advisory Committee establishes subcommittee on how SBE can promote robust research practices
- Summer & Fall 2013
  - Subcommittee proposal for workshop on "Robust Research in the Social, Behavioral, and Economic Sciences."
- February 20-21, 2014
  - Workshop convened
  - Participants drawn from variety of universities, funding agencies, scientific associations, and journals
  - Cover broad range of issues from extent and cause of problems to possible solutions
  - Details are in the appendix of the circulated report
- Post-workshop period
  - Document & digest workshop content
  - Discuss and propose recommendations
  - Complete report

Subcommittee Report

- DEFINITIONS
  - No consensus in science on the meanings of replication, robustness, generalizability
  - Different terms can refer to same thing
  - Same term can refer to different things
- RECOMMENDATIONS
- CONCLUSIONS

DEFINITIONS

- REPRODUCIBILITY
  - refers to the ability of a researcher to duplicate the results of a prior study using the same materials and procedures as were used by the original investigator
    - E.g., a researcher uses the same raw data, builds same analysis files, and same statistical procedures to make sure that same results obtained as in published study
    - Differences could be due to:
      - Processing (e.g., treatment missing data) of data
      - Application of statistical method (e.g., different defaults)
      - Accidental errors in original analysis (or follow-up analysis)
  - Reproducibility is a minimum necessary condition for a finding to be believable and informative.

DEFINITIONS

- REPLICABILITY
  - refers to the ability of a researcher to duplicate the results of a prior study if the same procedures are followed but new data are collected
    - a failure to replicate occurs when one study documents relations and a subsequent attempt with new data fails to yield the same relations
    - Null results or nonzero results could be replications
      - E.g., failure to find intervention to work in two different data sets is a replication as would be the finding of positive effect
    - Same researcher performing second study more likely to replicate
      - Fully aware of procedures
    - Second researcher in another location less likely because:
      - Did not directly observe the first study
      - Relies on text description of first study
      - Critical details not fully understood or described
      - Failure to replicate might be due to different procedures

DEFINITIONS

- GENERALIZABILITY
  - refers to whether the results of a study apply in other contexts or populations that differ from the original one
    - degree to which found relations apply in different situations
    - E.g., do findings based on college students apply to adult population of the United States?
    - E.g., does an experiment that uses one type of persuasive message work when researcher tries other types of persuasive messages?
    - Failure to generalize directs attention to operation of limiting conditions on relationship
      - Chance to advance theory as these limiting conditions are uncovered

DEFINITIONS

- REPRODUCIBILITY
  - refers to the ability of a researcher to duplicate the results of a prior study using the same materials and procedures as were used by the original investigator
- REPLICABILITY
  - refers to the ability of a researcher to duplicate the results of a prior study if the same procedures are followed but new data are collected
- GENERALIZABILITY
  - refers to whether the results of a study apply in other contexts or populations that differ from the original one

RECOMMENDATIONS

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

    - Ideally materials used to collect, transform, and analyze data would be archived in publically accessible online storage facility
    - If confidentiality issues preclude sharing, summary statistics or other means could be investigated to encourage reanalysis

RECOMMENDATIONS

- REPLICABILITY

- ***Recommendation 2***: *NSF should sponsor research that evaluates various approaches to determining whether a finding replicates and to assess which approach(es) under which circumstances are the most helpful for reaching valid conclusions about replicability.*

  - *Meaning of replicability seems intuitive but becomes murkier the closer we look*
  - *Does replication require same coefficient value and same statistical significance?*
  - *Can we say result replicated when sign and statistical significance are the same?*

RECOMMENDATIONS

- REPLICABILITY
- ***Recommendation 3***: *To permit assessing replication in various ways, NSF should encourage researchers to report associations between variables using different metrics (e.g., standardized and unstandardized coefficients, effect sizes, odds ratios) and indicating precision of estimates (with standard errors) and to assess the statistical significance of findings using these different methods.*
  - *We do not know the best way to assess replications, so best to have studies report effects in different metrics and estimates of variability*

RECOMMENDATIONS

- GENERALIZABILITY
- ***Recommendation 4***: *NSF should sponsor research that identifies optimal procedures for practically assessing all types of generalizability of findings (e.g., from a set of study participants to a population, from one set of measures to other measures, from one set of circumstances to other circumstances) and differentiating lack of generalizability from failure to replicate.*
    - *Failure to find same results using different measures, time points, or populations suggest important interactions or moderator variables*
        - *Generates theoretical questions on scope of effects*

RECOMMENDATIONS

- STANDARDS FOR STATISTICAL REPORTING
- ***Recommendation 5***: *NSF should fund research exploring the optimal and minimum standards for reporting statistical results so as to permit useful meta-analyses.*
  - *Small N can create difficulties*
    - *Low statistical power to detect effects*
      - *"No significant difference" in replication due to low power*
    - *Low cost experiment with small N might be run repeatedly and failures to replicate ones discarded because of "implementation issues"*
      - *Bias introduced in that only replications preserved*
      - *Low cost of small N experiment makes it more likely*
    - *High cost experiment (e.g., fMRI) might restrict N*
  - *Meta-analysis across different studies could aggregate results*
    - *Requires sufficient details on the results from contributing studies*
    - *Need guidelines on optimal & minimum reporting standards*

RECOMMENDATIONS

- STEPS TO DETER CONFIRMATION BIAS
  - Confirmation bias is the tendency to search for or interpret information in a way that confirms one's preconceptions or hypotheses, to avoid exposure to challenging information, and to discredit the challenging information one does encounter.
  - Questionable practices:
    - Fail to report results for all measures, but only those that worked
    - Collect more data if original data fails to support hypotheses
    - Only report experimental conditions favorable to expectations
    - Stop collecting data once desired result obtained
    - Excluding data points that undermine results
    - Keeping data points on which favorable results depend
    - Reporting unexpected results as if hypothesized in advance

RECOMMENDATIONS

- STEPS TO DETER CONFIRMATION BIAS
  - Confirmation bias is the tendency to search for or interpret information in a way that confirms one's preconceptions or hypotheses, to avoid exposure to challenging information, and to discredit the challenging information one does encounter.
- ***Recommendation 6**: NSF should support research into the use of questionable research practices, the causes that encourage such behavior, and the effectiveness of proposed interventions intended to discourage such behavior and should support the identification of empirically-validated optimal research practices to avoid the production of illusory findings.*

RECOMMENDATIONS

- STEPS TO DETER CONFIRMATION BIAS
  - Confirmation bias is the tendency to search for or interpret information in a way that confirms one's preconceptions or hypotheses, to avoid exposure to challenging information, and to discredit the challenging information one does encounter.

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

RECOMMENDATIONS

- UNDERSTANDING SCIENTIFIC PRACTICE
  - Conventional practices might unintentionally create robustness issues
  - Ideal design might be weakened by actual practice
    - Participants might not comply with treatment
    - Attrition of subjects
    - Errors in measurement
  - SBE Scientists are ideally suited to study human behavior & attitudes and their impacts on science in practice
    - In-depth interviews, participant observation, + other qualitative methods
    - Surveys, administrative data, meta-analysis permit other angles on problems
    - Can be applied to SBE and other science areas of NSF

RECOMMENDATIONS

- UNDERSTANDING SCIENTIFIC PRACTICE
  - ***Recommendation 8***: *NSF should sponsor research seeking to document suboptimal practices that are widespread in particular fields, with an eye towards identifying those areas that most depart from the scientific ideals and contribute to nonrobust research findings.*

RECOMMENDATIONS

- CONTINUITY IN MONITORING & MODIFICATION

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

FINAL COMMENTS

- Robust research is issue throughout sciences represented at NSF
- Scientists in the SBE disciplines have numerous tools to enable us to better understand the issues
- We could study SBE research and research in other sciences
- The subcommittee report provides our recommendations
- We welcome input