Tech Mining Cited References to Understand the Influence of Journal Articles on Reports of the US National Research Council

Jan Youtie, Georgia Tech USA
Barry Bozeman, Andrew Kao, Arizona State USA
Sahra Jabbehdari, Mulesoft

Motivation

- Contribution: little empirical understanding of use of STI
  - Much literature on use of formal information in decision-making (Simon 1944, 1991)
  - No literature on use of STI in science, technology and innovation (S&T) policy (Hammond et al., 1983, Bozeman et al, 1978)
- Research questions for study
  - Does the perception of the limited use of formal scientific and technical information (STI) accord with empirical reality?
  - How credible is STI compared with other sources?

Definition of STI

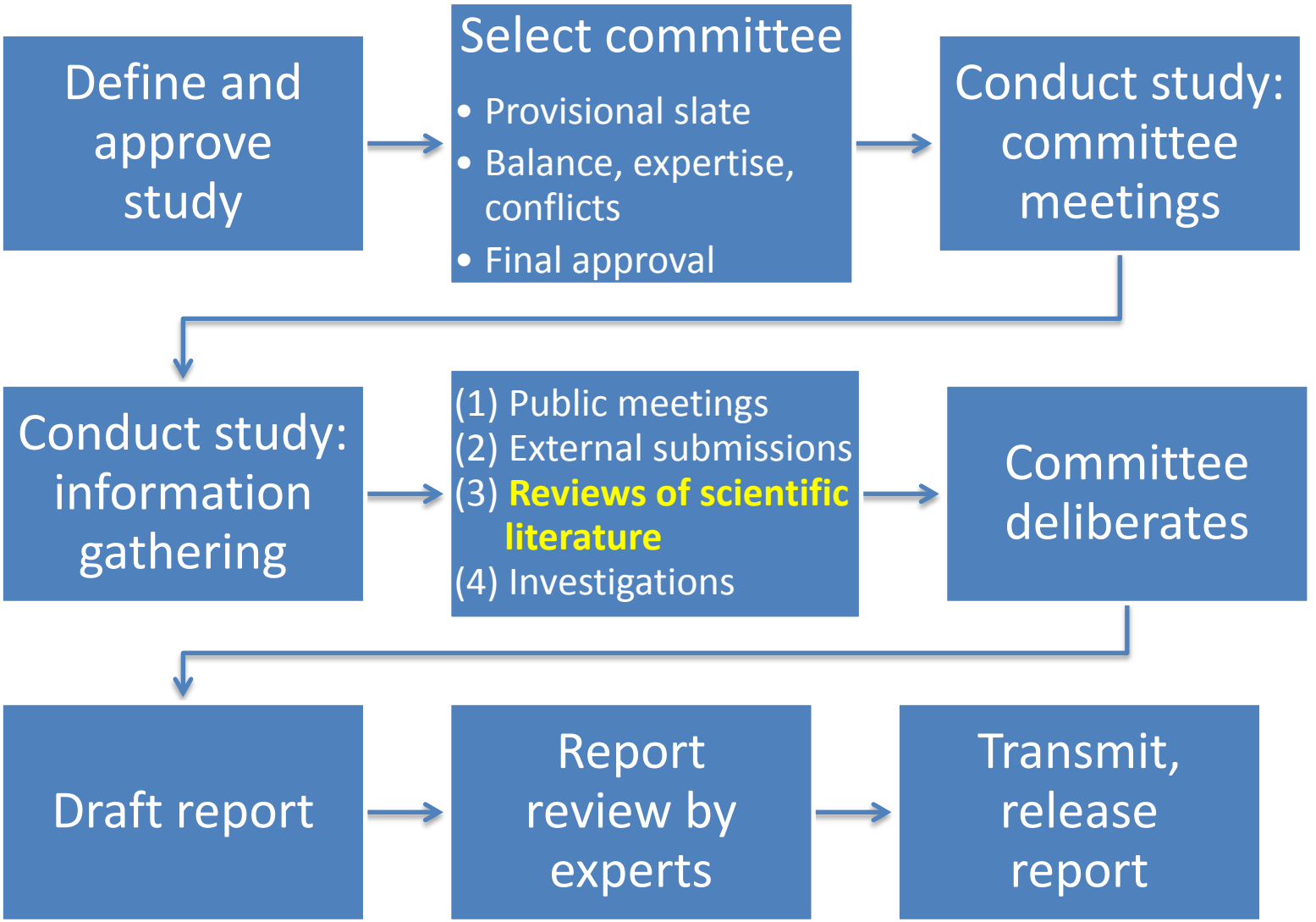
- ***Open scientific and technical literature appearing in peer-reviewed academic journals or proceedings.***
- STI used in a narrow sense v. typical in the literature (McClure, 1988; Walker and Hurt, 1990)

National Research Council (NRC) – an STI intermediary?

- Performs research work for the production of reports on science and technology issues within the National Academies
  - National Academy of Science: 1863
  - National Academy of Engineering: 1964
  - Institute of Medicine: 1970
- National Academies serves as advisor about science and technology intensive policy issues to Congress
- Little research on the NRC (despite $224M federal budget FY2014)*
  - Ellefson (2000): single case on non-forest federal land management
  - Policansky (1999): anecdotal (ecologist, NRC staff, well-constructed committees with high level of trust given a precisely constructed policy question are most successful)
  - Shapiro and Guston (2006): discursive (bureaucracies will shirk their duties, relying on the peer review process for correction)
  - Fein (2011): case (NRC plays an increasingly important role in regulatory peer review)
  - Parascandola (2007): history (conflict of interest policy in the NAS and NRC)
  - Martin and Irvine (1989): discursive (lack of priority setting in NRC reports)

*National Academies 2014 Report to Congress.

National Research Council Process



National Academies (2006) Our Study Process: Ensuring Independent, Objective Advice.

STI= *f*(Report Characteristics
(length, year published, policy area, NRC references Congressional authorization)
Committee Characteristics
(chair committee, reviewer sector)**)**

Sample

- All National Academies reports published 2005-2012
- Focus on board appointed/empaneled single shot studies (mostly NAS)
  - Exclusion of workshops
  - Repeated Congressionally authorized standing studies, i.e., Transportation Bureau (NAE), Health and Safety (IOM)
- Results=589 reports

Method

- NRC annual reports
  - Text-mine and code the NRC Reports, including cited references (footnotes v. listing)
    - Very time-consuming – took about a year
- Database linkages with
  - Web of Science
  - Scopus

Variables

- Report variables
  - Publication year
  - Policy area
    - Defense
    - Education
    - Tech transfer/industry
    - Environment,
    - Science
  - # pages (logged)
  - Congressionally authorized

Variables

- Committee variables
  - # committee members
  - Committee members by sector (business, government, academia, other)
  - Sector of chair
  - Reviewers by sector

Variables

- Reference variables:
  - # cited references
  - % STI (journal article, published proceeding) / # cited references
  - % other NRC reports / # cited references

Collecting and Cleaning STI

- Automated cleaning and matching methods (using Excel macros) to a thesaurus of journal articles
- Manual checking and coding of the references, by two separate coders, to determine whether or not the references were STI.
- Some NRC reports had a separate list of cited references
- Fewer than 20% of the reports used a footnote convention rather than a list of cited references at the end of the report.
  - Defense policy area report commonly used footnotes in lieu of a list of cited references.
  - Manually extract the footnotes before coding

Example of footnotes in NRC reports

[20]R.T. Drmanac and R.B. Crkvenjakov, Hyseq Technology. April 13, 1993. Method of sequencing genomes by hybridization of oligonucleotide probes. U.S. Patent 5,202,231. R.T. Drmanac and R.B. Crkvenjakov, Hyseq Technology. February 20, 1996. Method of determining an ordered sequence of subfragments of a nucleic acid fragment by hybridization of oligonucleotide particles. U.S. Patent 5,492, 806. R.T. Drmanac and R.B. Crkvenjakov, Hyseq Technology. June 11, 1996. Method of sequencing by oligonucleotide probes. U.S. Patent 5,525,464. R.T. Drmanac and R.B. Crkvenjakov, Hyseq Technology. September 16, 1997. Method of sequencing genomes by hybridization of oligonucleotide probes. U.S. Patent 5,667,972. R.T. Drmanac and R.B. Crkvenjakov, Hyseq Technology. December 9, 1997. **J. Eggers, K.M. Beattie, J. Shumaker, M. Hogan, R. Varma, J. Lamture, M.A. Hollis, D. Ehrlich, and D. Rathman. 1993. Genosensor technology. Clinical Chemistry 39:719-722. S.P.A. Fodor. 1997. DNA sequencing—Massively parallel genomics. Science 277:393. E.M. Southern. 1982. Application of DNA analysis to mapping the human genome. Cytogenet. Cell Genet. 32:52-57. E.M. Southern. 1982. New methods for analyzing DNA make genetics simpler. Biochemistry Society 10:1-4.** Method of sequencing by hybridization of oligonucleotide probes. U.S. Patent 5,695,940. J. Baier, Hyseq Technology. March 16, 1999. Reagent transfer device. U.S. Patent 5,882,930. R.T. Drmanac and R.B. Crkvenjakov, Hyseq Technology. October 26, 1999. Computer-aided analysis system for sequencing by hybridization. U.S. Patent 5,972,619. R.T. Drmanac and R.B. Crkvenjakov, Hyseq Technology. January 25, 2000. Method of sequencing genomes by hybridization of oligonucleotide probes. U.S. Patent 6,018,041. R. Drmanac, Hyseq Technology. February 15, 2000. Methods and apparatus for DNA sequencing and DNA identification. U.S. Patent 6,025,136.

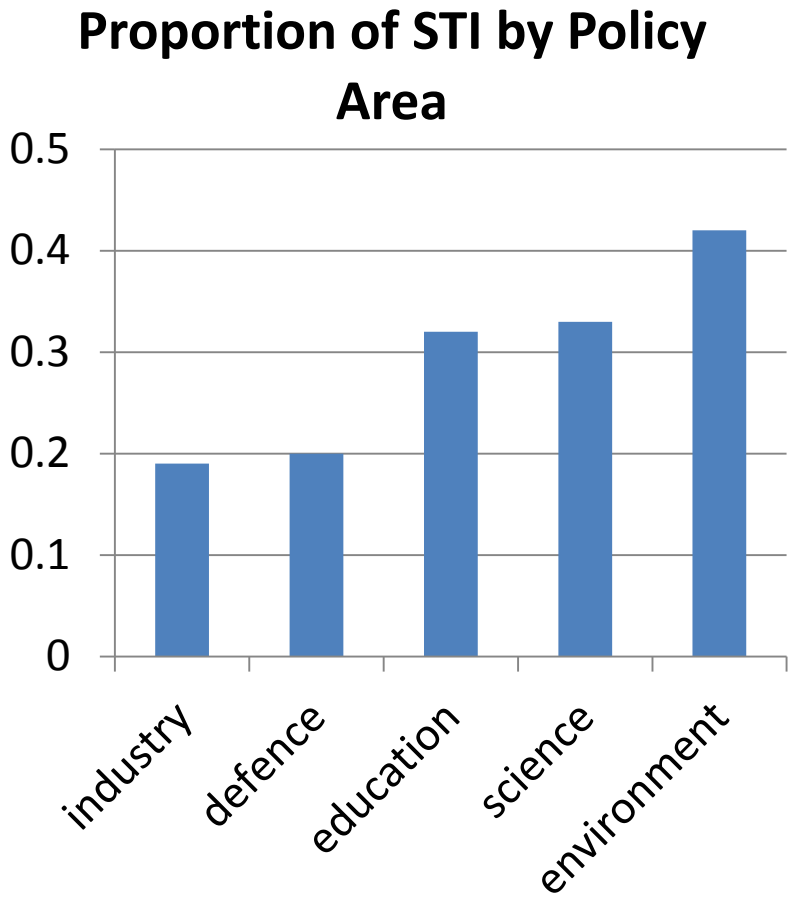
[41]Lizardi et al., 1998. See note 39 above.

[1]For references, see p. 37.

[32]Bacillus spores are very widely distributed in the environment.

Descriptive Statistics

- Reports size:
  - Range 16→650 (median=164, mean=188)
- 120,000+ references
  - in all but 3 reports
- STI=88% of reports,
  - Range: 0→1,440 (median=30, mean=89)
  - Simple STI/references (median=.26, mean=.30)
  - STI/pages preserves outliers



OLS Results
(Dependent Variable=STI)

- ↑ # pages (logged)
- ↑ Year published
- ↑ Environment
- ↓ Education
- ↓ % private sector committee
- ↓ % private sector reviewers
- ↓ Congressionally authorized
- ↓ % other NRC reports

STI Use Typology

SCIENTIFIC: "The committee typically places 'more faith in' articles from top journals, experiments with proper design and methods, and highly cited articles that were proven to be accepted by the scientific community. In addition, they looked for articles with larger sample sizes, and compared different experimental quality (controlled studies were more used than observational ones)."

DECENTRALIZED: "Each person on the committee was in charge of examining the literature in his or her field of expertise." / "Authors of individual chapter decide what references are used."

JUSTIFYING: "STI is used to support what they saying. In fact, they did not find literature disagree at anything. The literature was supportive of everything. Only in one case, they cited on both sides of the disagreements." / "The reviewers asked STI to be included more: primarily add citations and justify particular comments." / "Most of the time, these anecdotes/conversations were then looked up and searched for STI that backed this information up."

HARMONIZING: "They performed literature searches and relied on members' knowledge and judgment that "defined the field." They wanted to 'reach consensus on what evidence shows.'"

ANECDOTAL: "Anecdotes were used as illustration."

EXPERIENTIAL: "This effort was not about collaborations or specific research topics derived from STI…rather an assessment based on information gathered by the committee from outside speakers, almost all from the government, during its meetings and from our own collective knowledge." / "Experience of committee members and non committee members hugely important."

Conclusions

- STI widely used in NRC reports
  - Use varies by policy area
  - Use varies by sector
  - And by report characteristics: size, year, authorization
- STI is used different ways by different committees
  - Not just as scientific evidence
  - STI not as credible as experience or anecdotes in some committees
- Limitations
  - Not able to rank STI versus other information sources
  - Not able to extend to role of academics in science policy

Acknowledgements

- This work was supported by the US National Science Foundation, Science of Science and Innovation Policy, Award #1262251. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the sponsors.

- For more information

  http://stip.gatech.edu/credibility-and-use-of-scientific-and-technical-information-in-science-policy-making/