THE UNIVERSITY OF CHICAGO

JUDGMENTS OF SCIENTIFIC QUALITY AND THEIR EFFECTS ON PUBLISHED KNOWLEDGE AND ITS DIFFUSION

A DISSERTATION SUBMITTED TO

THE FACULTY OF THE DIVISION OF THE SOCIAL SCIENCES

IN CANDIDACY FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

DEPARTMENT OF SOCIOLOGY

BY

MIKHAIL TEPLITSKIY

CHICAGO, ILLINOIS

JUNE 2016

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would like to thank the members of my dissertation committee, James Evans (chair), Andrew Abbott, and Karin Knorr, for their guidance and encouragement. I would also like to thank John Levi Martin for invaluable advice over the years. Most important thanks are, as always to my mom Faina, sister Jane, and friends who supported me throughout this long journey. My father Abram passed away before this dissertation was finished, but he was in my mind through every page.

ABSTRACT

Collaborative efforts like modern scientific research depend on methods to evaluate and absorb participants' contributions, and at the research frontier this evaluative step is often accomplished through the peer review of grants and manuscripts. With billions of dollars and space in prestigious journals hinging on the decisions of reviewers, the review system has attracted consistent scrutiny. Many of the thousands of studies scrutinizing peer review focus on the reliability, validity, and fairness of the reviewers' decisions. Largely absent in this debate about peer review's internal practices are the consequences of these practices for the character and diffusion of published knowledge. This dissertation shifts the focus to the consequences of peer review practices through four case studies. The first case investigates the negotiation of revisions authors of quantitative sociological manuscripts undertake during peer review and reveals that substantial changes concern primarily manuscripts' theoretical framing, while the data analyses remain relatively stable. The case argues that the greater relative value placed on data and analysis over frames incentivizes investment into the former over the latter. The second case interrogates the common practice of using post-publication citations to evaluate the validity of review decisions. Analysis of the reviews of manuscripts submitted to the *American Sociological Review* from 1977 to 1981 and the manuscripts' subsequent citations reveals no relationship. However, reviewers' comments show that reviewers focused on the soundness of the manuscripts' arguments, not their potential impact. The case shows that a review process that results in publications of variable impact is not necessarily a failing of peer review, but rather a consequence of reviewers and citers draw on different dimensions of value. The third case study examines the consequences for quantitative sociology of the common bias for positive findings in peer review. Using hundreds of studies that use the *General Social Survey*, the published

statistical relationships are perturbed by slight changes to the model specifications. Results show that at the time of publication, results are relatively robust to this perturbation. Additionally, the published relationships are estimated using waves of the Survey that appeared after publication. Results indicate that published findings are weakened much more by social change. The last case focuses on the consequences of scientific peer review judgments outside of the sphere of science. By measuring rates at which millions of scientific journals are used as sources in Wikipedia, the largest online encyclopedia, I show that Wikipedia editors preferentially use high impact and the more accessible (open access) journals. The case shows that increased accessibility of the scientific literature improves its diffusion to the lay public and that a status ordering that review practices establish in one sphere, science, may be exported wholesale to a disparate context, Wikipedia.

CHAPTER 1. INTRODUCTION

Modern science is an enormous collaborative effort that depends on formal and informal institutions to evaluate individuals' contributions and link them to rewards. Although the primacy of evaluative institutions, particularly the various instantiations of peer review, is clear, their effectiveness is widely debated: the scholarly literature alone includes articles and books that number in the thousands[1]. Much of this literature questions the reliability, validity, and fairness of review decisions. It is consistently found that reviewers of grants and manuscripts disagree frequently (Bornmann, 2011a; Cicchetti, 1991a), with levels of agreement often "comparable to rates found for Rorschach inkblot tests" (Lee, Sugimoto, Zhang, & Cronin, 2013). The findings concerning validity and fairness of review decisions are much less consistent. Lacking the benefits of experimental control[2], most analysts attempt instead to measure associations between review decisions and proxies of quality, such as citations, or between decisions and authors' social and institutional characteristics (Danthi, Wu, Shi, & Lauer, 2014; Li & Agha, 2015; Siler, Lee, & Bero, 2015). Associations of review decisions with proxies of quality are generally small to nil, while associations with social or institutional characteristics vary substantially (Lee et al., 2013).

What is less clear than the consistency of findings from the reliability and validity and fairness literatures is their epistemological or normative import. Are disagreements between reviewers problematic and, if so, can disagreement be minimized, and by how much? Is a small or null association between review outcomes and citations problematic? Are the higher manuscript acceptance rates of senior scientists problematic? At the heart of these questions is a

[1] For instance, the Eugene Garfield collection at the University of Pennsylvania Library indexes more than 3700 publications. http://garfield.library.upenn.edu/histcomp/peer-review_to/. Accessed 2016-03-22.
[2] Important exceptions include (Boudreau, Guinan, Lakhani, & Riedl, 2016; Mahoney, 1977; Rooyen, Godlee, Evans, Black, & Smith, 1999).

concern with how reviewers arrive at their decisions, and the consequences that particular review practices have for the character and direction of the resulting science. Establishing the causes of peer review decisions has proven tremendously difficult – it is telling that, even after thousands of studies, the concerns that animate the peer review literature closely resemble those outlined by Zuckerman and Merton more than 40 years ago (Zuckerman & Merton, 1971). Quantitative work has by and large shirked taking into account the rich sociocultural contexts of review decisions (Sabaj Meruane, González Vergara, & Pina-Stranger, 2016). Meanwhile, qualitative work that takes richness of the context seriously has only recently attempted to link variation in context with consequences (Huutoniemi, 2012; Michèle Lamont, 2009a). Moreover, existing work has focused almost exclusively on the natural sciences, while evaluative practices in the social sciences have remained relatively unexamined (Camic, Gross, & Lamont, 2011; Leahey, 2008a). This dissertation presents four case studies oriented to the consequences of review practices, particularly in social science. Before describing and contextualizing the dissertation's case studies, I review the primary conceptual frameworks used to analyze review decisions.

Perspectives on review decisions

Classical philosophy of science and the Mertonian tradition

Philosophers have traditionally separated the problems of scientific discovery into two categories: the context of discovery and the context of justification (Kuhn, 1977b, p. 326). The context of discovery concerned the genesis of scientific ideas within scientists' minds. Being a matter of psychological and sociocultural processes, it fell within the purview of the behavioral sciences. On the other hand, the context of justification concerned the choice among proposed scientific claims. Being a matter of evidence and logic, it was the proper purview of philosophy. Philosophers of this tradition sought to develop a decision algorithm on which a scientist could

rely to choose the best among a set of theories. Although the particulars of the algorithm were debated, there was some consensus about what the algorithm should take into account, including a theory's accuracy or scope (Kuhn, 1977b, pp. 321–2).

The pioneering studies of evaluation in science by Robert Merton and his colleagues (R. K. Merton, 1968; Zuckerman & Merton, 1971) trace their conceptual underpinnings to this philosophical tradition. The Mertonian line of research was concerned primarily with whether evaluations in science were fair and consistent – concerns that have continued to animate the literature to this day. A fair decision was viewed as one that was uncorrupted by social, or "particularistic" factors. Fair evaluations should privilege the cognitive characteristics of articles and grants, such as their accuracy or scope, over the social characteristics of their authors, such as their academic rank (Lamont 2009: 18). Because there was in principle one decision algorithm to be used to assess a claim's true quality, reviewers should, in principle, be able to reach consensus regarding this value. Lamont and Mallard (2005) identify this Mertonian tradition, in which "scientific quality" had a fairly stable meaning across reviewers, as the first wave of evaluation studies.

The second wave of evaluation studies

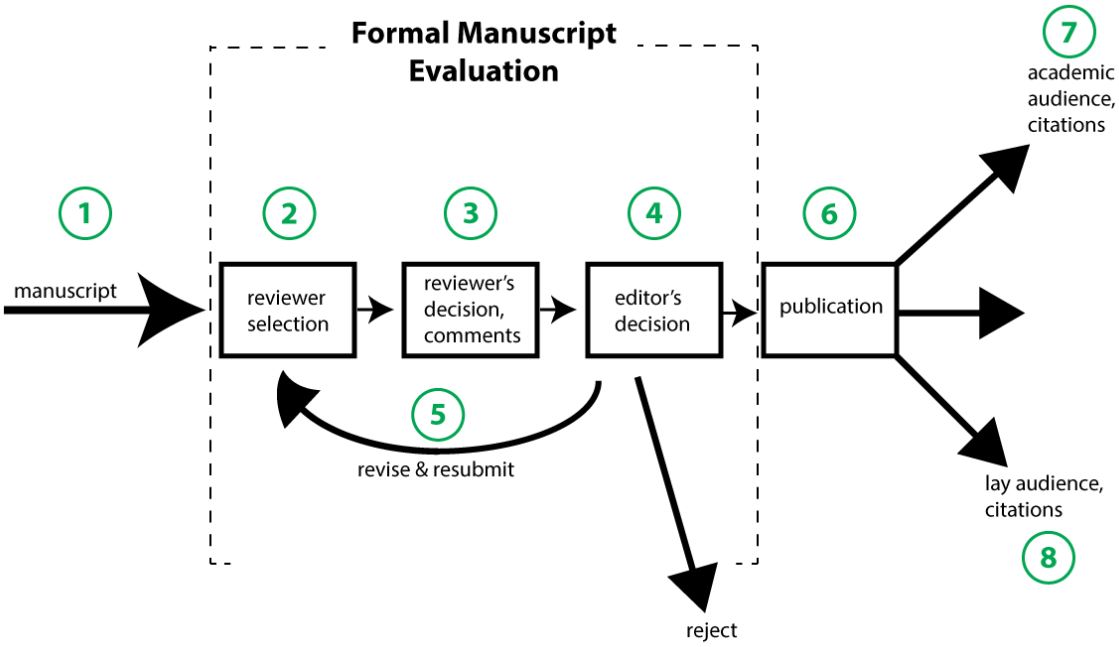
Yet already in 1973, Thomas Kuhn considered the search for a universal evaluation algorithm a dead-end: "Most philosophers of science would … now regard the sort of algorithm which has traditionally been sought as a not quite attainable ideal." (Kuhn, 1977b, p. 326). Moreover, Kuhn argued that epistemic criteria are vague enough as to allow normatively unproblematic variation in their application (326). Equally unproblematic were differences in how scientists weight the distinct criteria. What Lamont and Mallard (2005) call the second wave or evaluation research takes Kuhn's observations as the starting point and focuses on how

scientists deploy abstract evaluative criteria in concrete settings. Studies in this tradition, almost exclusively qualitative, emphasize interpretive flexibility of evaluative criteria. They eschew the traditional normative concerns and conceptual categories of the philosophy of science and explore how definitions of quality are debated, constrained, and negotiated (Camic et al., 2011; Joshua Guetzkow, Lamont, & Mallard, 2004; Hirschauer, 2010; Huutoniemi, 2012; Michèle Lamont, 2009a; Langfeldt, 2001; Mallard, Lamont, & Guetzkow, 2009). Work in this tradition has produced conceptual development and more nuanced descriptions of evaluation in practice, but it has, by and large, failed to link variation of practices with their consequences for knowledge (exceptions include Huutoniemi, 2012; Langfeldt, 2001). Yet the subject of consequences is fertile for exploration, and may be the likeliest bridge between the second wave of evaluation studies and the normative concerns of the first. Unlike the problem of identifying epistemologically problematic or neutral causes of individual reviewers' decisions, which is contentious even conceptually, it may be easier to make plausible normative claims about the consequences of particular review schemes. For instance, a peer review system that consistently publishes claims that fail to reproduce or are factually incorrect would likely be seen by scholars of all persuasions as normatively undesirable (J. P. Ioannidis, 2005; Open Science Collaboration, 2015). Linking particular review practices to their consequences for knowledge is the focus of this dissertation.

Case studies of peer review and its consequences

This dissertation consists of four case studies, primarily in the field of sociology. Each of these studies concerns one aspect of the overall process of formal evaluation, which is illustrated in a stylized fashion in Figure 1.1.

Figure 1.1. Illustration of the steps typically involved in manuscript review and publication. Each step is numbered for the purposes of identifying the focus of the dissertation's case studies.



The Figure displays the steps many, but not all, academic journals use to review submitted manuscripts. Authors send manuscripts to the journal (1), the editor(s) of which select one or more reviewers (2). The reviewers provide their evaluations (3) to the editor(s), who then decide(s) whether to reject, seek revisions (5), or publish the article (6). Once published, the article becomes available for reading by a number of audiences, including an academic audience (7), some members of which may cite the article, and a more general audience (8), some members of which may also cite the article, for example in Wikipedia.

The dissertation's first case concerns steps (1) and (5): it takes a sample of 30 published quantitative sociological articles and compares them to earlier drafts in order to measure changes the authors undertook during review. The results show that in cases where the connection between theoretical framing and data analysis is found wanting, the negotiation between authors

and reviewers results in substantial changes primarily to the framing, while the data analysis remain relatively stable. The asymmetry in the revisions reveals that data and analysis are relatively "expensive," that is valued relatively higher than theoretical frames, which can be, and are, substituted readily. The Conclusion argues that this review practice incentivizes the production of manuscripts in which authors invest more heavily in data than theory.

The second case concerns steps (3) and (7): it interrogates the line of research that tests the validity of peer review decisions by comparing them to alleged proxies of quality, i.e. citations. Using the reviews of manuscripts submitted to the *American Sociological Review* from 1977 to 1981 and the manuscripts' subsequent citation impact, I show that the relationship between decisions and citations is nil. In contrast to similar studies that have interpreted the absence of a relationship as a whole-sale indictment of peer review, I focus on the content of the review reports to show that reviewers did not seek to predict impact. Instead, reviewers evaluated the manuscripts' soundness, and it is judgments of soundness that are decoupled from subsequent impact. This finding turns the attention again to how reviewers interpret evaluative criteria.

The third case concerns steps (4) and (8): it focuses on judgments of scientific quality in the context of Wikipedia, the world's largest online encyclopedia. By measuring rates at which millions of scientific journals are used as sources, I show that Wikipedia editors preferentially use high impact journals, but make more use of the more accessible (open access) literature. The significance of this case is in the first instance pragmatic. The results document for the first time the effects for the general public of making scientific knowledge more accessible. Additionally, the case shows how the quality ordering of scientific claims conventional to one sphere – science – is exported wholesale to a somewhat disparate sphere, Wikipedia, despite differences in the missions of the two spheres.

The last case concerns steps (4) and (6): it examines the consequences for quantitative sociology of the prevalent bias for positive findings in peer review, which I take to exist but do not measure directly. Concern over this bias has been long-standing (Sterling, 1959) and many believe it makes the published literature a repository of false findings (J. P. Ioannidis, 2005). In this case study I seek to measure the consequences of this publication bias empirically. Using hundreds of studies that use the *General Social Survey*, the published statistical relationships are perturbed by slight changes to the model specifications. Results show that at the time of publication, results are relatively robust to this perturbation. Additionally, the published relationships are estimated using waves of the Survey that appeared after publication. Results indicate that published findings are weakened much more by social change.

CHAPTER 2. FRAME SEARCH AND RE-SEARCH: HOW QUANTITATIVE SOCIOLOGICAL ARTICLES CHANGE DURING PEER REVIEW*

Abstract

Peer review is a central institution in academic publishing, yet its processes and effects on research remain opaque. Empirical studies have (1) been rare because data on the peer review process are generally unavailable, and (2) conceptualized peer review as gate-keepers who either accept or reject a manuscript, overlooking peer review's role in *constructing* articles. This study uses a unique data resource to study how sociological manuscripts change during peer review. Authors of published sociological research often present earlier versions of that research at annual meetings of the American Sociological Association (ASA). Many of these annual meetings papers are publicly available online and tend to be uploaded before undergoing formal peer review. A data sample is constructed by linking these papers to the respective versions published between 2006 and 2012 in two peer-reviewed journals, *American Sociological Review* and *Social Forces*. Quantitative and qualitative analyses examine changes across article versions, paying special attention to how elements of data analysis and theory in the ASA versions change. Results show that manuscripts tend to change more substantially in their theoretical framing than in the data analyses. The finding suggests that a chief effect of peer review in quantitative sociology is to prompt authors to adjust their theoretical framing, a mode or review I call "data-driven." The data-driven mode of review problematizes the vision of sociological research as addressing theoretically motivated questions.

* Originally published in the journal *The American Sociologist* (Teplitskiy, 2015a) and reprinted here with permission from Wiley.

Introduction

Peer review, despite its limitations[1] continues to play a central role in scholarly publishing as scientists rely on this method of evaluation to maintain the integrity of the published literature and to distribute rewards (Harnad, 2000). Research on the outcomes of peer review is plentiful, but little remains known about its processes, especially in the social sciences[2]. Researchers can speculate about how peer review "works" using their own referee reports and outcomes, and perhaps the experiences of talkative colleagues, but aggregate patterns are unclear. Yet it is in the social sciences, where the published journal article is often an iterative product of authors, reviewers, and editors, that peer review may be most crucial; here it transcends its function of quality control to perform constructive work.

One aspect of manuscripts submitted to journals is of special interest: the connection between theory[3], data, and method, the last two of which I henceforth call "data analysis." A tight coupling between theory and data analysis is not easy to accomplish in social science, where many measures struggle against a variety of ambiguities (Abbott, 1997) and even the meaning of the word "theory" varies (Abend 2006, 2008; Rueschemeyer 2009: Ch. 1). Yet sociologists value the connection between theory and data analysis highly. For example, in her study of grant panels, Michele Lamont (2009: 182) found that nearly three-quarters of the panelists mentioned that the connection between theory and data analysis was an important ingredient of good grant proposals. What happens when this connection is lacking? In grant peer

[1] Critiques of peer review are plentiful. Useful reviews include (Bornmann, 2011b; Campanario, 1998a, 1998b; Michèle Lamont, 2009a; Lee, Sugimoto, Zhang, & Cronin, 2013).

[2] Recently a literature has started to coalesce around research practices, including evaluation, in the social sciences. See (Camic, Gross, & Lamont, 2011, p. 200; Michèle Lamont, 2009a; Leahey, 2008a) for reviews and examples. Leahey (2008: 45) identifies the processes of framing papers and the receptiveness of authors to criticisms – the subjects of this study – as particularly ripe for investigation.

[3] Later sections of this document will use the concept of a "theoretical frame," rather than "theory." The distinction is developed later in the introduction.

review, the outcome may be rejection of the proposal. But what about journal peer review, where reviewers who find that the data analysis does not fit the theory can make recommendations to improve the fit?

It is useful to consider this question in light of two ideal-types of review: question-driven and data-driven. In the question-driven mode the commitment to a specific theory-relevant research question is central. If journal reviewers find that the data analysis an article performs to answer a research question is unsuitable, they expect the author to try to obtain different data and perform a different analysis, but the research question guiding her research will remain constant. In the data-driven mode research questions are guided by the availability of data and possibility of analysis. When reviewers find a misfit between theory and data analysis, they expect the author to produce a more suitable theoretical frame, keeping the data analysis as-is. What mode of review is prevalent in sociology? The nature of the constructive role played by peer review has been hitherto unclear owing largely to the secrecy of the process.

The present study attempts to overcome the traditional secrecy of peer review by utilizing a unique data resource. Authors of published sociological research articles sometimes note if they had previously presented the research at an Annual Meeting of the American Sociological Association (ASA). Many of these annual meeting papers are publicly available on the Internet and have not been peer reviewed by a journal. I link a sample of these papers to their subsequently published versions in two peer-reviewed journals, *American Sociological Review* and *Social Forces*. Each article pair captures the development of a research project at two points in time. It is reasonable to assume that papers at the first point in time, presentation at ASA, often contain the same key theoretical and data analytic elements as the manuscripts submitted to

journals and subsequently published[4]. It is also likely that peer review, perhaps consisting of several rounds at one or more journals, is *a* central force provoking authors to make revisions across versions[5]. Changes in papers from ASA to publication are thus indicative of the effects of peer review on manuscripts[6].

Qualitative and quantitative analyses of the article pairs concentrate on changes in data analyses and theoretical frames across versions. The chief finding is that the amount and nature of the data analysis in an article, operationalized as the set of variables used, tends to increase and change slightly from the ASA paper to the published article. On the other hand, the amount and nature of (explicit) theoretical framing, operationalized as the set of references from the article's theory and literature review sections, tends to increase and change substantially. This finding suggests that a chief achievement of peer review may be to provoke authors to adjust their theoretical framing while leaving the bulk of the data analysis intact. This "data-driven" mode of review problematizes the vision of sociological research as addressing theoretically motivated questions[7].

This study contributes to past research on peer review in both conceptual and pragmatic ways. Conceptually, past research has been preoccupied with *grant* peer review and the natural sciences. The chief activity of grant peer review is quality *evaluation* (and prediction). Since the pioneering studies of Merton and colleagues, scholarship has concentrated on whether quality evaluation – the "yes" or "no" decision – is fair, consistent, and, to a lesser extent, discipline-

[4] One of the limitations of this study, elaborated later, is that it does not examine manuscripts that were ultimately rejected but only those that were ultimately published.
[5] To be more precise, this study assumes that the changes provoked by journal peer review do not differ systematically from changes effected by other forces, e.g. informal comments in workshops and from colleagues.
[6] This approximation is further justified in the "Analytic strategy" section.
[7] Note, for example, that many ASA sections are topical. The discipline's bibliometric structure also appears problem-centered (Moody and Light 2006).

dependent (see Bornmann 2011, Lee et al. 2013 for thorough reviews). The literature on the rhetoric of science has also focused on the natural sciences, especially its devices of persuasion, such as the impersonal tone, and the way in which published articles misrepresent actual research (Gilbert 1976; Knorr and Knorr 1978; Knorr 1977; Gilbert and Mulkay 1984; Latour and Woolgar 1979; Dear 1985; Shapin 1984).

A shift in attention to the social sciences raises unique questions (Camic et al., 2011; Leahey, 2008a). Journal peer review, especially in sociology, is most often an interactive and collective process, one which is as much about quality *construction* as its evaluation. Additionally, sociological theory is in many areas not developed to a point of generating concrete predictions (Cole 2001; Rueschemeyer 2009: 1). The theory that is found in the literature thus often comes in the form of "frames," which may be defined as objects of theory more specified than sociology's grand perspectives (e.g. symbolic interactionism) that appear in introductory textbooks (Lynch & Bogen, 1997) but less so than theories that generate specific falsifiable predictions (Rueschemeyer 2009: 1). The present article focuses specifically on the constructive aspect of review and on theoretical frames.

It is also important to understand the practical consequences of particular review practices. The avalanche of publications is making it increasingly difficult to recruit overworked reviewers to a system many find problematic. It is thus important to take stock of how the current system functions in practice so that its desirable aspects may persist and others may be abandoned. In particular, it is important to understand the incentives that various forms (or modes) of review give to researchers. For example, it is possible that different modes of review incentivize researchers to invest more heavily in theory or new data or data analysis.

The rest of this paper is organized in the following way. The next section discusses previous studies of peer review, which by and large have studied what predicts whether a manuscript is accepted or rejected. The following section elaborates the analytic strategy by laying out more explicitly its assumptions. Next, I discuss the corpus of article pairs and the qualitative and quantitative analyses used to analyze them. I then discuss key findings and conclude with a discussion of the results, alternative interpretations, and implications for the development of sociological theory.

Lastly, it should be emphasized that this paper seeks to understand the effects of the peer review *process* and not its outcomes in particular cases, certainly not to append normative labels to particular cases. The specific article pairs examined below appeared in top sociological journals and thus represent some of the discipline's best work.

Literature review

Peer review

A broad review of the peer review literature may be found in the introduction. In the context of this case study it is important to note that much of this literature, and certainly the studies within it that have garnered the most attention, examined *grant* peer review. Stephen Cole and colleagues (1981) studied NSF grant panels, Wenneras and Wold (1997) examined a Swedish grant competition and Michele Lamont (2009) studied social sciences and humanities grant panels. Perhaps the biggest conceptual difference between grant and journal peer review is that the outcome of grant peer review is typically a "yes" or a "no," while the outcome of journal peer review is typically a "no" or "revise and resubmit" (Bakanic et al. 1987: 634), *both of which come with reviewers suggestions for how the manuscript may be improved*. Journal peer review is thus in part about the *construction* of quality, not only its evaluation. Furthermore, the second

wave of peer review research has emphasized the interaction between reviewers in the evaluation of quality. But in journal peer review, especially in "revise and resubmit" cases, the interaction between the author(s) of the manuscript and editors and reviewers is most prominent. To date, this alternate type of interaction has been underexplored (exceptions include Goodman, Berlin, Fletcher, & Fletcher, 1994; Gross, 1990; Myers, 1985; Simon, Bakanic, & McPhail, 1986). The present study takes as its subject the interactive *creation* of quality of (sociological) journal articles, rather than evaluation of quality of grant proposals.

Rhetoric of science

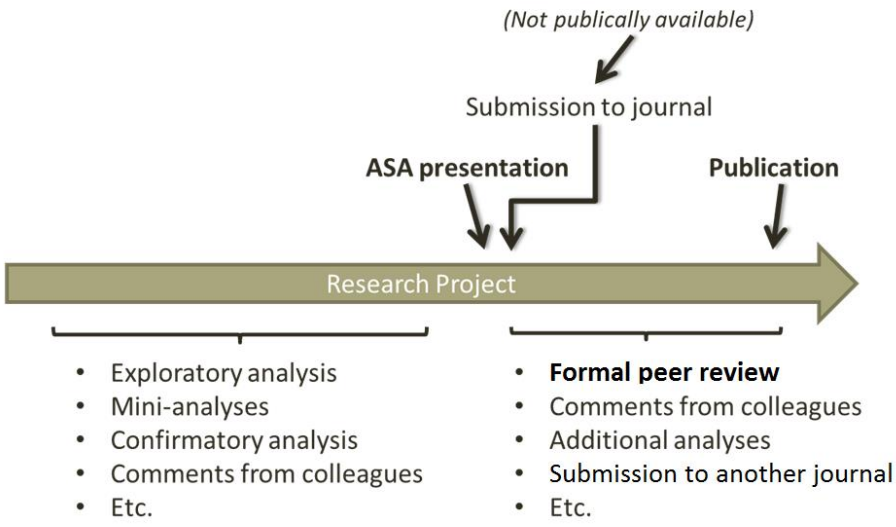
Another literature to which this study contributes concerns the rhetoric of science. Science writing, like any other writing intended to persuade, is rhetorical, and much has been written about how authors (wittingly or unwittingly) use persuasive literary devices, especially in the natural sciences (Bazerman 1983, Gross 1990a; Bazerman 1988). These include strategic deployment of the circumstances of production (Gilbert and Mulkay 1984), adding or deleting modalities (Knorr-Cetina 1981; Latour and Woolgar 1979), tone (Gusfield, 1976), references (Gilbert, 1977), and audience management (Lamont 1987; Shapin 1984). Another group of studies has compared actual research practices to the published reports and found that the latter to be grossly inaccurate summaries of the former (Gilbert 1976; Knorr and Knorr 1978; Medawar 1964). By and large these works have viewed publication as a step in which some elements of research are strategically unmentioned and others packaged into an article intended for an audience. The present study, on the other hand, emphasizes that the process by which research is made public often consists of a step in which a small collective composed of reviewers and editors negotiate the research elements (Bakanic, McPhail, and Simon 1989; Myers 1985) and may contribute elements entirely absent from the original research. Secondly,

this study shows that outcomes of the negotiations commonly result in changes to theoretical framing.

Analytic strategy

The analytic strategy of this study is based on a stylized chronology of quantitative research projects in sociology displayed in Figure 2.1 below.

Figure 2.1. A stylized trajectory of quantitative sociological research. Formal peer review may include submission/rejection/resubmission from several journals. The key data for investigating peer review, the manuscripts submitted to journals, are generally unavailable.

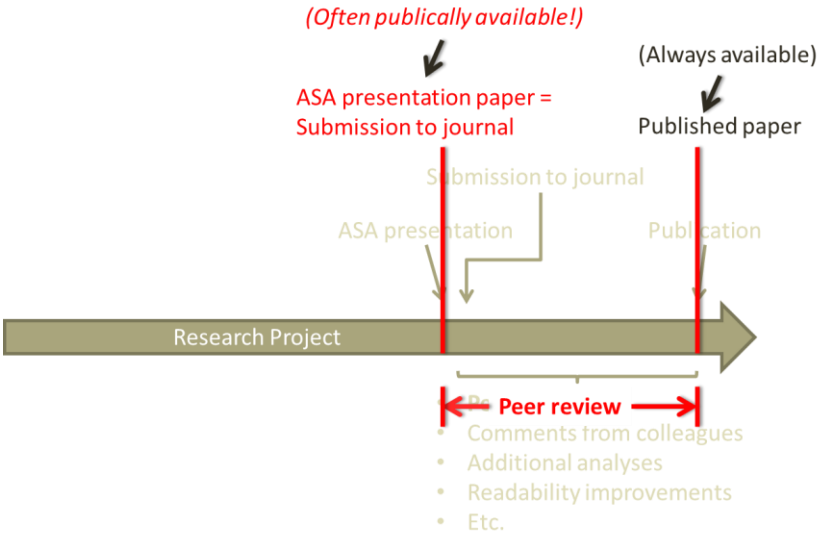


The thick arrow tracks the passage of time from left to right and is divided roughly into two phases. The first phase of sociological research is highly nonlinear – hypothesis-generation and data analyses occur simultaneously (Abbott 2004: Ch. 7; Abbott 2014: 4-6; Glaser and Strauss 1999). In this phase data is tinkered with (Knorr, 1979), a research question is tentatively formulated, tested, reframed, tested again, and so on, all the while the researcher reads relevant literature, and receives comments from colleagues, workshop audiences, and so on. Eventually the project may assume the shape of a draft submitted to an annual meeting of the American

Sociological Association (ASA). It is reasonable to assume that this ASA draft includes the central elements of both data analysis and theory present in the manuscript the author submits to a journal. As the author's goals converge on publication, additional data gathering and analysis, literature review, and theoretical reframing take a back seat as the manuscript undergoes journal peer review (Abbott 2014: 5 [Figure 1]). Once reviews are received, the author may make revisions and, ideally, the manuscript is published, which is where the thick arrow stops.

Peer review in this stylized multi-step process is difficult to study due to data availability. The correspondence between author, editor, and reviewers is an established secret in academics. The outcomes of this correspondence may nevertheless be approximated by making two simplifying assumptions, pictured in Figure 2.2.

Figure 2.2. A stylized trajectory of quantitative sociological research with two key assumptions: ASA manuscripts often contain the same elements of data analysis and theory as manuscripts submitted to journals, and in the period between submission and publication, manuscripts change in response to forces of which peer review is central. *These assumptions enable comparison between article versions to observe how peer review affected elements of the original (ASA) paper.*



First, it is reasonable to assume that papers at the first point in time, presentation at ASA, often contain the same key theoretical and data analytic elements as the manuscripts submitted to journals and subsequently published. The data analyses presented below do not presume that the manuscript versions are identical and, in fact, the different word limits of ASA and published papers generally preclude perfect equivalence; it is only expected that the key elements of theory and data analysis that are present in the ASA papers are also present in those submitted to journals. This assumption is sufficient because the data analyses concentrate on the fate from ASA to publication of these "original" elements only. For example, the analyses do not take into account if the published version of a paper contains many more references because, perhaps, there is now "space" for them; they only take into account *changes* made to references *present* in the ASA version (and, presumably, the manuscript submitted to the journal).

Second, it is likely that peer review is a central force provoking authors to make changes across versions. Of course changes, if any, from the earlier (ASA) to the later (published) version of an article may come about for a variety of reasons: journal peer review, comments from colleagues, further work by the researcher, etc., but it is not clear why changes stemming from these different sources would differ systematically from one another. For example, it is not clear why comments from colleagues would tend to provoke an author to edit a paper's framing but comments from reviewers would tend to provoke the author to edit data analysis.

These assumptions enable the identification of the effects of peer review upon a sample of sociological articles. I concentrate on how peer review affects one aspect of articles – the connection between theory and data. Reviewers weigh the connection between theory and data very heavily when judging research. Michele Lamont found that nearly three quarters of social scientists on the panels she studied ranked this aspect of articles as important (2009: 182). She

writes, "Panelists wax poetic – evoking a language of beauty and appreciation – in describing proposals that reach perfect articulation between the research question, the theory informing the research, the method proposed, and the evidence mobilized to answer the question" (182). Following Lamont's terminology I divide the four features of the articulation above into two groups:

> *theoretical framing*: the research question and the theory informing it
> *data analysis*: data, methods and the evidence mobilized to answer the question.

To extract the theoretical framing and data analysis from articles I rely on their conventional division into five main sections: introduction, literature review and theory, data and methods, results, and discussion/conclusion. I identify *theoretical framing* with the literature review and theory sections, and *data analysis* with the data and methods, and results sections.

What happens when reviewers find that the connection between theoretical framing and data analysis is imperfect? The author(s) may respond by changing either nothing, the framing, the analysis, or both. Do authors overwhelmingly respond in one way or another? Authors' typical responses signal two different modes of journal peer review. In the *question-driven mode*, a researcher is centrally concerned with answering a particular question and the data analyses are chosen for their fruitfulness in reaching an answer. If peer reviewers find that the fit between the theoretical framing and the data analysis is unsuitable, the researcher will respond by performing a different data analysis, one that will hopefully better fit to the theoretical framing. In this mode of research, the chief change to an article between ASA and final publication would be in the data analysis. In this mode, the data source, model or set of variables tends to change, while the theoretical framing remains relatively constant.

In the *data-driven mode*, a researcher who is interacting with peer reviewers is committed to a set of data and method of analysis. If peer reviewers find that the fit between the theoretical framing and the data analysis is unsuitable, the researcher will respond by changing the theoretical framing to one that will hopefully be better supported/tested by the data analysis. In this mode the theoretical framing tends to change, while the data analysis remains relatively constant.

Data

This study analyzed thirty pairs of articles, where each pair consists of (a) an unpublished article presented at an annual meeting of the American Sociological Association, and (b) the version of that article which was eventually published in either *American Sociological Review* (*ASR*) or *Social Forces* (*SF*), two of the most central American sociology journals. To collect these pairs, I searched online databases of *ASR* and *SF* for the keywords "American Sociological Association," and "meeting" or "meetings." Phrases such as these often appear in the notes and acknowledgments sections of articles to indicate that the research had been previously presented at annual meetings of the ASA. I then attempted to locate the previous versions of the articles on the ASA website, which maintains a database of presentation papers from 2003 to present. For each journal, I sorted the search results by publication date and took, beginning in June 2012 and working backwards, the 15 most recent articles that were quantitative methodologically and whose ASA versions were available for download.

The sample was limited to quantitative articles for a pragmatic rather than conceptual reason: tracking changes in the variables and data used in different versions of quantitative articles proved to be relatively straightforward and unambiguous. For articles published in *ASR*, I examined 37 articles before finding a suitable 15. 12 of the 37 lacked records in the ASA

database, 3 had records but had not been uploaded, and 7 were qualitative methodologically. For articles published in *SF*, I examined 23 articles before finding a suitable 15. 5 of the 23 lacked records in the ASA database, 2 had records but had not been uploaded, and 1 was qualitative methodologically. Table A2.1 in the Appendix presents the pairs of articles used.

Methods

Qualitative

The pairs of articles from *ASR* and *SF* were analyzed both qualitatively and quantitatively. In the qualitative analysis, I closely read both versions of the articles, paying attention to all changes from the title to the references. It became clear during this work that some sections of articles change substantially, while others remain relatively constant (e.g. sections describing the data used). I subsequently concentrated on the way the author(s) framed her thesis within existing theory and the data analysis she performed to evaluate the. In the next section I present numerous examples and descriptive statistics to capture the range of changes authors make to their theoretical framing and data analysis sections. I also attempt to describe changes quantitatively using two approaches, text similarity, and overlap in variables and references.

Quantitative

Text similarity

The close-reading of the article pairs led to the following question: Which sections of the ASA article look most like corresponding sections of the published article? To answer this question, I first divided each article into the conventional set of sections: introduction, theory and

literature review, data and methods, results, discussion and conclusion[8], and then divided each section into sentences using common text processing tools[9]. Figures, tables, and references were not included, as were stopwords [10] and punctuation. All words were converted to lower case. I also dropped from the calculations sentences of length less than 5 and characters less than 10 after finding that these short passages did not constitute true sentences or contributed too many false matches. For each sentences in each ASA section, I searched for best match among all the sentences of the published article[11].

The similarity between any two sentences A and B was defined as the Jaccard coefficient between the set of words in A and the set of words in B. The Jaccard coefficient of similarity is defined as:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}.$$

where A and B are sets. In the present case, A is the set of words in a sentence of the ASA article and B is the set of words in a sentence of the published article. This index takes on values from 0 (when A and B have no elements in common) to 1 (when A and B are identical). Two sentences that are identical or close matches are expected to have a large Jaccard coefficient. For example, consider sentences "*The coefficient of gender was not significant*." and "*The coefficient estimates of gender and race were not significant*." The Jaccard similarity between these sentences, after removing stopwords and other pre-processing, equals |{'coefficient', 'gender', 'significant'}| /

[8] Some articles, especially the ASA versions, lacked one or more of the following sections. Other articles, especially the published versions, included separate discussion and conclusion sections or an additional section providing context (e.g. historical). Nevertheless, all articles largely followed this five-section structure.
[9] I used Python 2.7 and the Natural Language Processing Toolkit 2.0.
[10] It is common in natural language processing to remove common words, often called stopwords, such as "the," "was," and the like.
[11] I also compared the sentences in an ASA article section to sentences only in the corresponding section of the published article, instead of its entirety. The results were qualitatively similar.

|{‘coefficient’, ‘gender’, ‘significant’, ‘estimates’, ‘race’}| = 3/5 = 0.8. I recorded the highest coefficient (best match) of an ASA sentence and took the average of these (highest) coefficients for all the sentences in a section. The resulting number is interpreted as “The sentences of section X in the ASA article have, on average, matches of quality Y in the published version.” This measure was designed to capture how well the text of each section in the ASA version is represented in the published version, and to disregard any *additional text* authors may have added to the published version.

In addition to text similarity between article versions, I investigated changes to theoretical framing and data analysis sections by measuring changes to the set of references used in the literature review/theory sections and the set of variables used in the data analysis, respectively.

Similarity of theoretical framing

The second way in which I measure changing theoretical framing uses changes to the *set of references* used in an article pair’s *theory and literature review sections*. The assumption behind this measure is that when theory informs the data analysis, such as a variable or model choice, *and* the *explicit* theoretical framing is considered necessary, the theory employed or contributed to will be documented in the references.

Treating the references in any section of the article as exact proxies for its theoretical components is problematic. The number of references in ASA articles is often limited by word count limits the document must conform to, while the references in published articles are influenced by journal (e.g. must cite publishing journal) and professional (e.g. self-citation) pressures. In an attempt to sidestep these distorting factors I concentrate on how many of the references that *were* made in the ASA article, in spite of the word count limits and often prior to

journal/professional pressures, were also made in the published paper. Specifically, I calculate the percent of these "old" references that appear in the published article.

Similarity of data analysis

To measure changes to the data analysis across article versions I record changes to the *set of variables* used in the analysis. The assumption behind this measure is that two article versions that use similar sets of variables are performing similar analyses. Following the similarity in theoretical framing measure above, I calculate the percent of variables in the ASA article that appear in the published article.

Several sources of ambiguity were encountered in coding variables from different article versions; all such situations were resolved by considering the variables distinct. In some cases categorical variables were often partitioned into different numbers of levels. In other cases continuous variables were occasionally disaggregated into several continuous variables of different types. Occasionally variables were intended to measure the same concept, but used different operationalizations and data sources. For example, in order to capture citizenship policies in different countries Christopher Bail used the number of naturalizations by region of former citizenship in the ASA paper (Bail 2006), but used a more direct measure (from a new data source) of these policies in the published paper (Bail, 2008). In all such cases, the corresponding variables in the article pair were coded as being different. Variables were considered the same only if they were entirely identical. Lastly, interaction terms that appeared in only one member of the article pair were considered unique variables.

Results

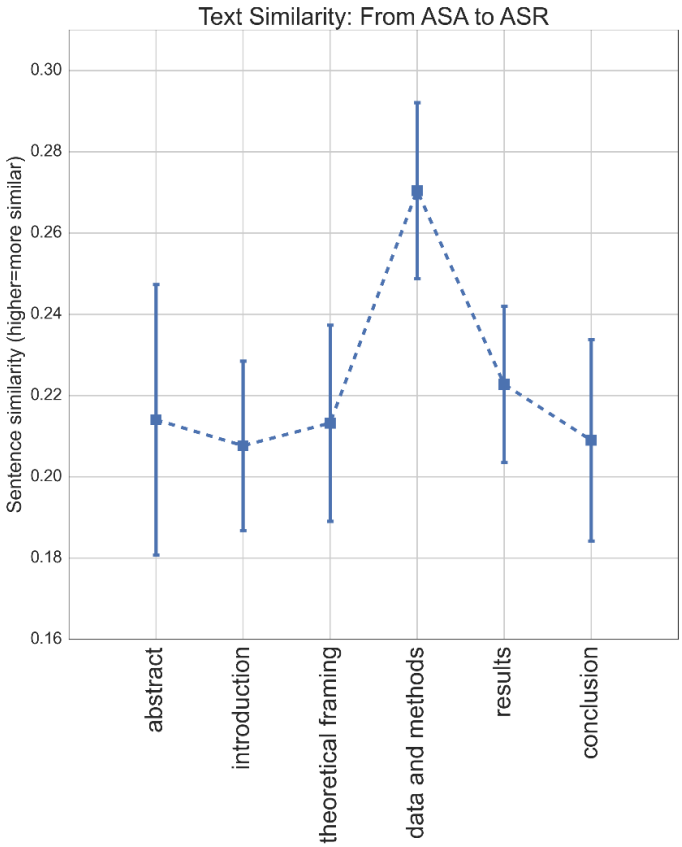
The chief finding of this paper is that while both theoretical framing and data analysis sections of sociology articles tend to change between ASA presentations and final publication,

theoretical frames change more substantially. This section is organized in the following way. First, I present results of the analysis of text similarity. Next I present results from measuring change using sets of references and variables used in data analysis, and illustrative examples of developmental trajectories that articles take. Lastly, the section examines whether changes in the data/analysis and theoretical framing tend to be related.

Text similarity

Figures 2.3 and 2.4 below present section-by-section text similarity between ASA articles and their published versions. Figure 2.3 was produced by averaging the text similarities of articles that were published in *ASR* while Figure 2.4 does the same for articles published in *Social Forces.* Both figures also include a dotted "baseline" curve, which is the average similarity between a section of an ASA paper and the same section of a *randomly chosen* published article. The baseline thus approximately represents similarities one expects between the sections of any two sociological articles.

Figure 2.3. Section-by-section text similarity between ASA papers and their *ASR* versions.



In Figure 2.3, the section in ASA papers that is most closely matched by sentences in the same section of *ASR* articles is the "data and methods" section. All other sections are substantially less similar to their published equivalents.

Figure 2.4. Section-by-section text similarity between ASA papers and their *Social Forces* versions.

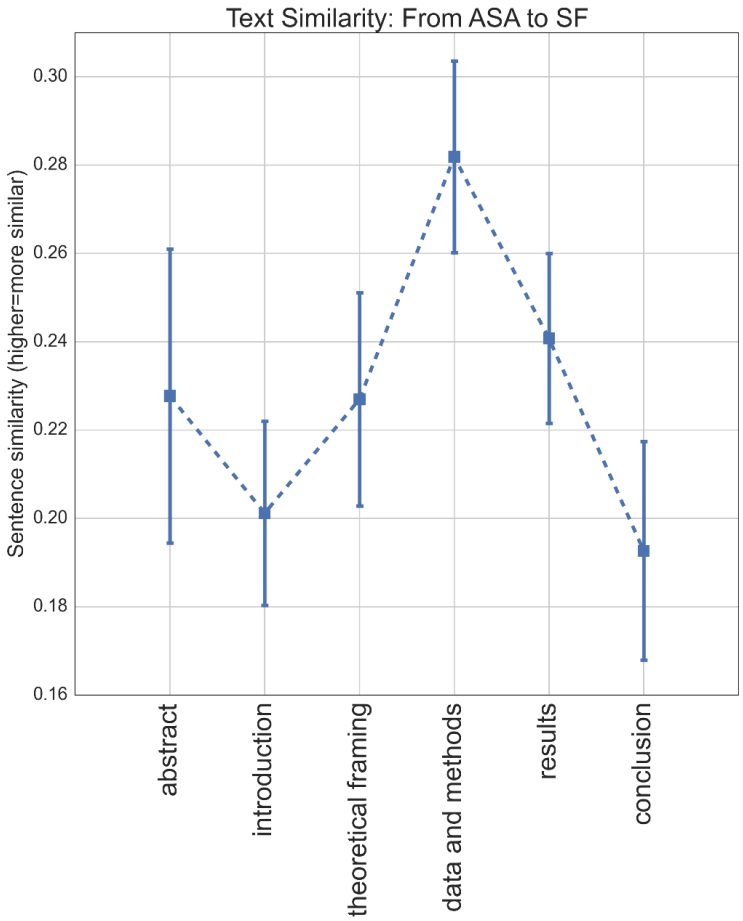


Figure 2.4 presents a qualitatively similar picture for articles published in *Social Forces*: "data and methods" section is the section least altered during peer review.

Next, I concentrate on changes that take place in data analysis and theoretical framing sections[12].

[12] The analysis of changes in variables and references uses a smaller number of article pairs (23) than the analysis of textual similarity (30) due to the increased difficulty of deep-reading the large number of pairs. The article pairs chosen for the more fine-grained analysis were those published most recently.

Changes in data analysis

For the purposes of this paper data and analysis are operationalized as data *sources* and *variables* used in models, respectively. Data is considered to change from one article version to another only if the author(s) uses a previously unused data source in the *published* version of the paper[13]. If a data source was used *only* in the ASA version, it is almost certain that the author was also aware of this data when editing the paper for publication but *chose* not to use it. The choice not to use this data source in the final version is most accurately described as an analysis choice, which ought to be captured in the different sets of variables used between the article versions.

Changes in the data

Using these definitions, a close reading of the article pairs in the corpus indicates that authors tend to use the same data sources in both versions of their papers. In 8 out of 10 article pairs, data sources did not change at all[14]. In one article pair (Armstrong et al. 2012), almost twice as much data became available (and was used) after the ASA paper presentation (Armstrong et al. 2009). The new data contained additional observations of both dependent and independent variables. In another other article pair in which data sources changed, an indirect measure of air pollution (Olzak and Soule 2007) was replaced in the final version of the paper with a more direct measure, using a new data source (Olzak and Soule 2009); in both analyses air pollution was a control variable. In summary, most articles in the corpus tend to use the exact same data in both versions of the paper.

[13] The emphasis on data source is designed to capture the intuition that when an author uses a large dataset, e.g. Current Population Survey, and uses some of the data in one version of a paper but other data in a different version, the data has not substantively changed, because the author was likely aware of all of the data within CPS the entire time.

[14] In one article pair, a data source was used in the ASA but not the final version of the paper.

Changes in the analysis

Analysis is considered to change only if the author(s) employs at least one variable new variable, for example in a regression equation[15], in either version of the paper. In the article pairs corpus, analysis changes were common. Tables 2.1 summarizes these changes.

Table 2.1. Numbers of variables used in ASA and published papers. The paired t-test does not reject the null hypothesis of no difference (p-value = 0.22)

Number of Variables		
	ASA version	Published version
Mean	23.3	25.5
Min	0	12
Max	77	88
Std	15.7	16.2

Published versions of articles use on average more variables (25.5) than the ASA versions (23.3). This does not mean that published articles on average use 25.5 variables in a regression model. Usually there are several regression models presented and each uses only a subset of the total variables used.

In order to assess how similar are the sets of variables used in different versions of articles, I computed the percentage of variables that are used in the ASA version that are also used in the published article. A high percentage indicates that the published paper *adds on* the earlier analysis; a low percentage indicates that the analysis has changed, rather than been added on to. The percent of "old" variables used in the "new" version of a paper is 58.9% (sd = 0.25).

[15] The distinction between data source and variable is not clear when the variable is used only descriptively (i.e. not in a regression equation). The number of such cases in the corpus is relatively small.

Changes in theoretical framing

The "theoretical framing" of articles was operationalized as the set of references *used in articles' literature review and theory sections*. Table 2.2 displays descriptive statistics of this measure in ASA and published articles.

Table 2.2. References in the literature review and theory sections of articles. The paired t-test does not reject the null hypothesis of no difference (p-value = 0.19)

Number of References		
	ASA version	Published version
Mean	34.4	40.5
Min	0	20
Max	73	85
Std	21.9	18.0

References in the literature review and theory sections change significantly between ASA and final versions of articles. First, the mean number of references increases (34 to 41). This increase is substantially larger than the increase in the number of variables used in analyses (Table 2.1). Secondly, the variability in reference usage is much greater than the variability in variable usage. The smallest number of references used in an ASA paper is 0. In essence, that paper[16] simply presented findings, without *explicitly* couching their motivations and significance in a sociological literature; the final version of the article contained a typical number of references. This case supports the hypothesis that explicit theoretical framing takes a secondary priority to data analysis for at least some authors.

In order to assess how similar are the sets of references used in different versions of articles, I computed the percentage of references used in the ASA version that are also used in the published article. A high percentage of overlap in references indicates that the published

[16] The authors indicated that they did not want the draft cited or circulated without permission.

paper uses a similar frame to its earlier version, although it may also *add on* to it; a low percentage indicates that the framing has *changed*. The percent of "old" references used in the "new" version of a paper is 31.0% (sd = 0.21). This similarity measure is lower than the corresponding measure of similarity in variables (analyses), and the means of the two samples are statistically different[17].

These quantitative results aggregate a number of different development trajectories of articles. The examples below illustrate some of these trajectories.

"Theoretical framing expanded, data same, analysis slightly different"

Cedric Herring's (2006) ASA paper investigated whether diversity in the workplace makes good business sense. The dimension of diversity in that article was racial diversity. In the *ASR* paper published in 2009, the argument remained the same, but to racial diversity was added a measure of gender diversity. The same data sources were used in both versions (though, curiously, the number of observations used in analyses was different even though the stated criteria for the selection of cases remained the same). The number of hypotheses tested doubled from four to eight: four for race and four for gender. The percentage of "old" *references* used in the "new" article is 0.33; the percentage of "old" *variables* used in the "new" article is 0.62.

"Theoretical framing expanded, data same, analysis same"

Christopher Bail's (2006) ASA paper examined the configuration of symbolic boundaries against immigrants in Europe. The version published in *ASR* in (2008) used the same data sources and presented the same analysis. The percentage of "old" *references* used in the "new" article is 0.21; the percentage of "old" *variables* used in the "new" article is 0.33. The large change in variables is due to recoding of substantively similar variables. The large change in

[17] T-test of independent samples: t = 3.95, p-value < 0.001

theoretical framing is due to the more than doubling of references used in the literature review and theoretical framing section of the published article (29 to 64). This example shows that while the substance of the paper has remained constant, the theoretical framing expanded.

"Theoretical framing different, data same, analysis expanded" – example 1

Dorius and McCarthy's (2006) ASA paper investigated the differential participation of leaders in voluntary organizations, paying special attention to gender differences. The published paper (2011) performed a similar analysis using the same data sources, but added a number of interaction terms. The percentage of "old" *references* used in the "new" article is 0.48; the percentage of "old" *variables* used in the "new" article is 0.77. The theory sections of these articles provide an interesting example of theory changes in theoretical framing that are not captured with crude quantitative measures. Hypothesis 5 of the ASA paper proposes that "**female leaders will volunteer fewer hours** per week than their male counterparts" (2006: 7, emphasis added). The data indicated otherwise, though this fact was not discussed in the discussion section. While the *Social Forces* (2011) version of the paper emphasized interaction effects, yet in Hypothesis 4 it too posited a direct effect for gender: "*The gender of the leader will have a direct impact on weekly leadership effort such* ***that female presidents will devote more time*** *each week to the organization than male presidents"* (2011: 459, emphasis added).

"Theoretical framing different, data same, analysis expanded" – example 2

Another example involving a change in theoretical framing and a corresponding expansion of analysis can be seen in the article pair Shorette and Hironaka (2010) and Shorette (2012). Both article versions attempt to "advance a theory of outcomes of world polity and to assess the environmental impact of cultural versus political economics factors in globalization processes" (2010: 2). The earlier ASA paper attempts to make its theoretical advance by evaluating the implications of three theories: (1) modernization theory, (2) political economy,

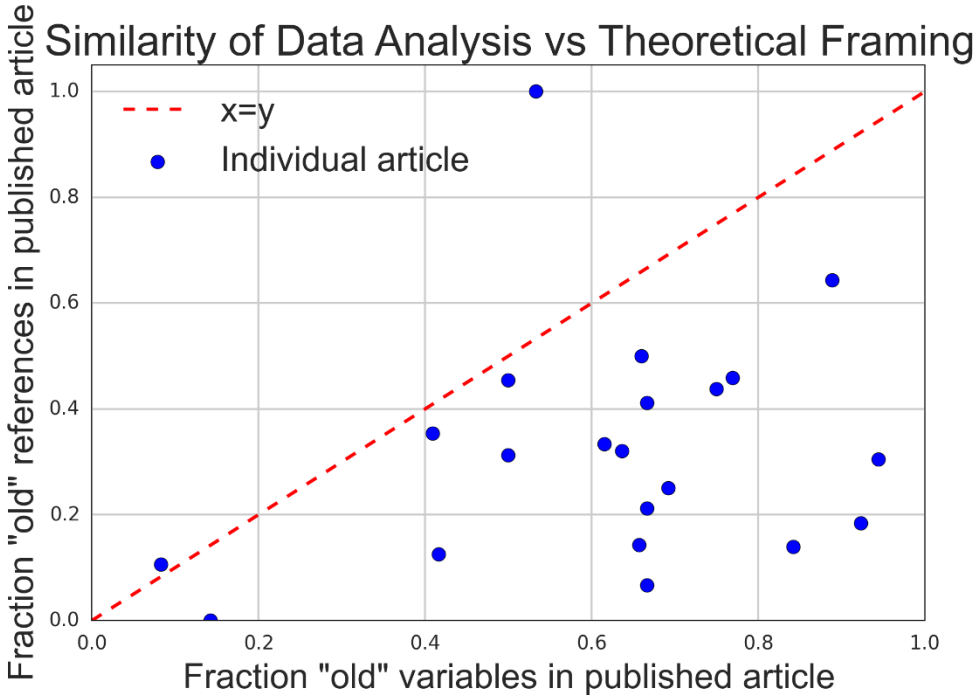
and (3) institutional theory. In the published version the comparison is between the implications of only two theories: (1) political economy and (2) institutional theory. The data remained the same. The analysis in the published version added a number of interaction variables and united two previously separate variables into a scale. The percentage of "old" *references* used in the "new" article is 0.42; the percentage of "old" *variables* used in the "new" article is 0.50.

Do changes in theoretical framing correlate with changes in data/analysis?

The results above indicate that between the ASA and publication, the theoretical framing section of articles tends to change more than data sources or data analysis. Are these changes correlated? A high correlation suggests a tight coupling between theoretical framing and data analysis. The coupling would likely be due to causality, potentially bi-directional: an additional analysis may signal that an existing theoretical framing is insufficient and thus requires change; or a previously unexplored theoretical consideration may prompt an additional analysis. A low correlation on the other hand would indicate a relatively loose fit between theoretical framing and data analysis.

Figure 2.5 presents a plot of the percentage of "old" (from the ASA paper) references used in the "new" (published) paper against the percentage of "old" variables used in the "new" paper.

Figure 2.5. Percent of variables from the ASA paper that are used in the data analysis in the published paper vs. percent of references from the theoretical framing section of the ASA paper that are used in that section of the published paper. The dotted line represents y = x, that is equal similarity between variables and references. Most of the points lie below the dotted line, which may indicate that in general there is more similarity in data analysis than theoretical framing.



The correlation coefficient is 0.24 (p-value = 0.29). Most of the points lie below the $y = x$ line (the dotted line), indicating that for a given article pair, the similarity in references is almost always lower than the similarity in variables. This pattern provides evidence for what I call the data-availability mode of peer review. In the data-availability mode, the data analysis remains relatively stable during peer review, while the author(s) and reviewers work out a suitable theoretical framing. In the question-centered mode, the theoretical framing remains relatively stable, as the author(s) and reviewers work out a suitable data analysis. In the question-centered mode, one expects points above the $y=x$ line, as these would indicate that the references are stable but the variables are not; there are few such points observed.

The findings presented above illustrate some common trajectories of how articles change in the time between ASA and publication. It should be emphasized again that these examples are not examples of poor sociological research but, rather, observations that help understand the pressures exerted by the peer review process. The following section emphasizes the shortcoming of the presented data analysis.

Conclusion

This article considered the understudied process of peer review in sociology. It focused on the effects of peer review on quantitative manuscripts. Several analyses demonstrated that between ASA annual meetings and publication, quantitative sociological articles change chiefly in their theoretical framing; data analyses change but not as markedly; data rarely change. There may be a positive correlation between changes in analysis and framing, but the sample is too small to speak to robustness. It was argued that ASA papers are likely to include key elements of data analysis and theory present in the manuscripts authors submit to journals for review. Consequently, changes in manuscripts from ASA to publication are likely to reflect revisions encouraged by reviewers and editors. Peer review in the sample of articles examined appears to operate in a data-driven mode, that is it treats theoretical frames as more change-able than data analyses. The relatively frequent and substantial variations to theoretical frames promoted by this mode of review problematize the vision of sociology as resolving theoretically motivated questions. In other words, there is a good deal of flexibility in how sociological analyses are interpreted, what literatures and theories they may speak to, or what their implications are. If peer review is indeed the major motivator of change, the character of research that emerges is one in which authors are either not aware of the most relevant theoretical frame or disagree with reviewers and editors what it is; the published theoretical frame thus appears to be a result of

negotiation between the authors, reviewers and editors, rather than a finely specified theoretical question that motivated the study in the first place.

Some ASA papers were composed almost entirely of the data analysis and included no references. There are two possible interpretations of sources of change in articles that present substantive findings at ASA and later couple these findings with a theoretical framing in time for publication. It is possible that the author had the theoretical framing in mind, but did not make it explicit in the ASA paper. Another possibility is that the framing was created later, perhaps at the request of reviewers. Both possibilities indicate that to be published in a sociology journal, an article must be theoretically framed. Furthermore, the latter possibility implies that a theoretical framing is necessary even if the theory did not motivate the research in the first place.

Limitations

This study suffers from several limitations. First and perhaps most crucially, it uses available data to make inferences about the unavailable peer review data. The assumptions entailed in the inferences have been discussed above. Regardless of the plausibility of the assumptions, it should be emphasized that articles can and do change, trivially or substantially, for a variety of reasons, peer review being just one. It is possible, if unlikely, that some of these forces produce changes systematically different from those produced by formal peer review.

Second, the data are missing those manuscripts which were presented at ASA, submitted to a journal, and rejected. While it is conceivable that the bulk of reviewers' criticisms of these ultimately rejected manuscripts concerned data analysis and not theoretical framing[18], the hypothesis is not supported by the available empirical data. For example, in Bakanic et al.'s

[18] For example, Fox (1989: 189) suggests that the disunity of sociology will push reviewers to comment on those matters over which there *is* consensus, namely methods and data anlysis.

(1987, 1989) study of peer review at *ASR*, reviewers made similar types of criticisms of accepted and rejected manuscripts, although the accepted manuscripts also earned praise; and in their study of the editorial process at *Administrative Science Quarterly* Strang and Siler (2015) find that the revision process focused primarily on conceptual and theoretical challenges.

Third, the measurement of theoretical framing via references is certainly crude. In many situations a certain idea can be attributed to a large number of different references. A specific reference that is chosen from a plausible set may be chosen to appease reviewers or the journal editor or due to a number of other considerations that have nothing to do with theoretical framing. In other words, references may simply be more fungible[19]. To account for this possibility at least in part, during the analysis of changes in sets of variables used between articles, variables were defined as being different using very loose criteria. For example, if a categorical variable in one version of an article was used as a number of dummy variables in the other version, all of these variables were coded as being different from each other. This deliberate coding choice was intended to make (conceptually) small changes in variables equally likely to be counted as (conceptually) small changes in references. Nevertheless, it is a major limitation of the present study that it cannot completely disqualify either of these interpretations.

Discussion

Sociologists, even in times of self-criticism, have devoted little attention to how peer review, and evaluation more broadly (Michèle Lamont, 2012), affects the development of theory or incentivizes various kind of effort and investment (Espeland and Sauder 2007; see Nosek, Spies, and Motyl 2012 for a discussion of incentives in psychology). For example, in a collection

[19] It should be noted that the fungibility of references or variables need not be taken as an exogenous fact; it may be taken, as it is in this study, as a phenomenon *to be* explained.

of critical essays entitled "What's Wrong With Sociology?," many of the distinguished contributors lamented the lack of progress in theory and almost as many diagnosed the ailment as due to sociologists choosing theoretical frames on ideological grounds (Cole 2001: 13). This study points out that theoretical frames may also be "chosen" collectively during the negotiation of peer review. Theory is also a product of investments into it; researchers must choose to invest into theory rather than other desirable goals. Literature must be found, read, and synthesized, its holes identified and filled. How different modes of review incentivize difference types of investments is an important question that is currently without a conclusive answer.

This study also brings attention to the concept of theoretical frames, which neither the sociology of science literature nor the rhetoric of science literature has developed fully. For example, in her review of social science research practices (Leahey 2008), Erin Leahey identifies theoretical framing as ripe for research, and does not find any relevant literature to cite. Theoretical frames, defined earlier as theories specified more fully than grand perspectives but not finely enough to generate specific falsifiable predictions, are commonly used by sociologists, who often speak about "framing" research this or that way. Frames may be objects unique to the social sciences and humanities. The present study demonstrated frames to be relatively changeable during revisions, implying that several frames may fit particular data and there may be disagreement about which one fits "best." Existing literature on the rhetoric of science has by and large concentrated on construction of the relatively specified theories and facts of the natural sciences (Gilbert and Mulkay 1984; Knorr-Cetina 1981; Latour and Woolgar 1979). The construction of theoretical frames is an analogous and appropriate subject for the growing literature on social science research practices and rhetoric (Camic et al., 2011; Leahey, 2008a; McCloskey, 1998; Nelson, 1990).

Lastly, it is valuable to study journal peer review in practice because, if developments in other fields are instructive, the traditional form of peer review will evolve. It will be up to sociologists which aspects of the traditional system persist and which will be abandoned. For example, observing the effects of revise-and-resubmit, as this study attempted to do, informs future decisions about the desirability of this practice.

Appendix 2

Table A2.1. Article pairs in the analyses. ASA paper which the authors requested not to be cited were used to compute aggregate statistics but not in finer grained analyses.

Data Sample	
ASA annual meeting paper	**Published Paper**
Armstrong, Elizabeth A., Paula England, and Alison C. K. Fogarty (2009). "Determinants of Women's Orgasms in College Hookups and Relationships."	Armstrong, Elizabeth A., Paula England, and Alison C. K. Fogarty (2012). "Accounting for Women's Orgasm and Sexual Enjoyment in College Hookups and Relationships." *American Sociological Review* 77(3): 435-462.
Brady, David, Andrew Fullerton, and Jennifer Moren-Cross (2007). "Putting Poverty in Political Context: A Multi-level Analysis of Working-aged Poverty across 18 Affluent Democracies."	David, Brady, Andrew S. Fullerton, and Jennifer Moren Cross (2009). "Putting Poverty in Political Context: A Multi-Level Analysis of Adult Poverty across 18 Affluent Democracies." *Social Forces* 88(1)*:* 271-300.
Buchmann, Claudia, Vincent J. Roscigno, and Dennis Condron (2006). "The Myth of Meritocracy? SAT Preparation, College Enrollment, Class and Race in the United States."	Buchmann, Claudia, Dennis J. Condron, and Vincent J. Roscigno (2010). "Shadow Education, American Style: Test Preparation, the SAT and College Enrollment." *Social Forces* 89(2): 435-62.
Budig, Michelle J., and Melissa J. Hodges (2008). "Differences in Disadvantage: How the Wage Penalty for Motherhood Varies Across Women's Earnings Distribution."	Budig, Michelle J., and Melissa J. Hodges (2010). "Differences in Disadvantage: Variation in the Motherhood Penalty across White Women's Earnings Distribution." *American Sociological Review* 75(5): 705-28.
Requested not to be cited	Castilla, Emilio J. (2011). "Bringing Managers Back In: Managerial Influences on Workplace Inequality." *American Sociological Review* 76(5): 667-94.
Dencker, John C. "Gender Differences in Career Trajectories: A Longitudinal Study of Promotion Patters in a Large US Firm."	Dencker, John C. (2008). "Corporate Restructuring and Sex Differences in Managerial Promotion." *American Sociological Review* 73: 455-76.

Table A2.1 continued	
Dorius, Cassandra and John D. McCarthy (2006). "The Differential Participation of Leaders in Nurturing a Social Movement."	Dorius, Cassandra and John D. McCarthy (2011). "Understanding Activist Leadership Effort in the Movement Opposing Drinking and Driving." *Social Forces* 90(2): 453-473.
Requested not to be cited	Fullerton, Andrew S., Wayne J. Villemez (2011). "Why Does the Spatial Agglomeration of Firms Benefit Workers? Examining the Role of Organizational Diversity in U.S. Industries and Labor Markets." *Social Forces* 89(4): 1145-64.
Requested not to be cited	Gibson, Christopher (2012). "Making Redistributive Direct Democracy Matter: Development and Women's Participation in the Gram Sabhas of Kerala, India." *American Sociological Review* 77(3): 409-34.
Goldstein, Adam (2010). "Revenge of the Managers: Labor Cost-Cutting and the Paradoxical Resurgence of Managerialism in the Shareholder Value Era, 1984-2001."	Goldstein, Adam (2012). "Revenge of the Managers: Labor Cost-Cutting and the Paradoxical Resurgence of Managerialism in the Shareholder Value Era, 1984 to 2001." *American Sociological Review* 77(2): 268-94.
Griffin, Larry J., and Kenneth A. Bollen (2005). "What Do These Memories Do? Civil Rights Remembrance and Racial Attitudes."	Griffin, Larry J., and Kenneth A. Bollen (2009). "What Do These Memories Do? Civil Rights Remembrance and Racial Attitudes." *American Sociological Review* 74: 594-614.
Harding, David (2006). "Disadvantaged Neighborhoods, Cultural Heterogeneity, and Adolescent Outcomes."	Harding, David (2009). "Collateral Consequences of Violence in Disadvantaged Neighborhoods." *Social Forces* 88(2): 757-84.
Herring, Cedric (2005). "Does Diversity Pay?: Racial Composition of Firms and the Business Case for Diversity."	Herring, Cedric (2009). "Does Diversity Pay?: Race, Gender, and the Business Case for Diversity." *American Sociological Review* 72(2): 208-224.
Hirsch, Elizabeth (2005). "Organizing Equal Employment Opportunity: The Effect of EEO Enforcement on Sex and Race Segregation in the Workplace."	Hirsch, C. Elizabeth (2009). "The Strength of Weak Enforcement: The Impact of Discrimination Charges, Legal Environments, and Organizational Conditions on Workplace Segregation." *American Sociological Review* 74: 245-71.
Isaac, Larry (2004). "Novel Countermovement Narratives: 'Fictions of the Real' as Cultures of Class in the Gilded Age."	Isaac, Larry (2009). "Movements, Aesthetics, and Markets in Literary Change: Making the American Labor Problem Novel." *American Sociological Review* 74: 938-65.

Table A2.1 continued	
Jackson, Margot I. (2007). "The Timing of Early-life Health and Socioeconomic Disadvantage."	Jackson, Margot I. (2010). "A Life Course Perspective on Child Health, Cognition and Occupational Skill Qualifications in Adulthood: Evidence from a British Cohort." *Social Forces* 89(1): 89-116.
Requested not to be cited	Janssen, Susanne, Giselinde Kuipers, and Marc Verboord (2008). "Cultural Globalization and Arts Journalism: The International Orientation of Arts and Culture Coverage in Dutch, French, German and U.S. Newspapers, 1955 to 2005." *American Sociological Review* 73: 719-40.
Kaya, Yunus (2006). "What Drives Industrialization in Developing Countries?: Globalization and Manufacturing Employment in 90 Developing Countries, 1980-2003."	Kaya, Yunus (2010). "Globalization and Industrialization in 64 Developing Countries, 1980-2003." *Social Forces* 88(3): 1153-82.
Kim, Hyojoung, and Steven Pfaff (2007). "Social Networks, Political Regime and Heterodoxy in the Reformation Movement."	Kim, Hyojoung, and Steven Pfaff (2012). "Structure and Dynamics of Religious Insurgency: Students and the Spread of the Reformation." *American Sociological Review* 77(2): 188-215.
Requested not to be cited	Kim, ChangHwan, and Christopher R. Tamborini (2012). "Do Survey Data Estimate Earnings Inequality Correctly? Measurement Errors Among Black and White Male Workers." *Social Forces* 90(4): 1157-81.
Thye, Shane, Edward J. Lawler, and Jeongkoo Yoon (2006). "Social Exchange and Micro Social Order: Comparing Four Forms of Exchange."	Lawler, Edward J., Shane R. Thye, and Jeongkoo Yoon (2008). "Social Exchange and Micro Social Order." *American Sociological Review* 73: 519-42.
Kalleberg, Arne L., and Ted Mouw (2006). "Occupations and the Structure of Wage Inequality in the United States, 1980s-2000s."	Mouw, Ted, and Arne L. Kalleberg (2010). "Occupations and the Structure of Wage Inequality in the United States, 1980s to 2000s." *American Sociological Review* 75(3): 402-31.
Olzak, Susan and Sarah A. Soule (2007). "Cross-Cutting Influences of Environmental Protest and Legislation."	Olzak, Susan and Sarah A. Soule (2009). "Cross-Cutting Influences of Environmental Protest and Legislation." *Social Forces* 88(1): 201-226.
Prechel, Harland, and Lu Zheng (2009). *No title*.	Prechel, Harland, and Lu Zheng (2012). "Corporate Characteristics, Political Embeddedness and Environmental Pollution by Large U.S. Corporations." *Social Forces* 90(3): 947-70.

Table A2.1 continued	
Requested not to be cited	Rosenfeld, Michael J., and Reuben J. Thomas (2012). "Searching for a Mate: The Rise of the Internet as a Social Intermediary." *American Sociological Review* 77(4): 523-47.
Shorette, Kristen and Ann Hironaka (2010). "Outcomes of World Polity: Trends in Chemical Fertilizer and Pesticide Consumption, 1961-2006."	Shorette, Kristen (2012). "Outcomes of Global Environmentalism: Longitudinal and Cross-National Trends in Chemical Fertilizer and Pesticide Use." *Social Forces* 91(1): 299-325.
Song, Lijun (2005). "When Institutions Meet Networks: Educational Homogamy in Urban China."	Song, Lijun (2009). "The Effect of the Cultural Revolution on Educational Homogamy in Ubran China." *Social Forces* 88(1): 257-70.
Requested not to be cited	Warren, John Robert and Caitlin Hamrock (2010). "The Effect of Minimum Wage Rates on Highschool Completion." *Social Forces* 88(3): 1379-1392.
Wolfinger, Nicholas H., Mary Ann Mason, and Marc Goulden (2006). "Dispelling the Pipeline Myth: Gender, Family Formation, and Alternative Trajectories in the Academic Life Course."	Wolfinger, Nicholas H., Mary Ann Mason, and Marc Goulden (2009). "Stay in the Game: Gender, Family Formation and Alternative Trajectories in the Academic Life Course." *Social Forces* 87(3): 1591-1621.
Yoon, Jeongkoo (2006). "Mechanisms Constructing Legitimacy of Team Supervisors and Their Effects on Team Efficacy and Team Commitments."	Yoon, Jeongkoo, and Shane Thye (2011). "A Theoretical Model and New Test of Managerial Legitimacy in Work Teams." *Social Forces* 90(2): 639-59.

CHAPTER 3. DO PEER REVIEWS PREDICT IMPACT? EVIDENCE FROM THE *AMERICAN SOCIOLOGICAL REVIEW*, 1978-1982[*]

Abstract

This study investigates how well peer reviews of articles published in the journal *American Sociological Review* between 1978 and 1982 predict the articles' citation impact in the following 32 years. We find no evidence of a relationship between review outcomes and citation impact at any time after publication, even when citations are normalized by subfield. Qualitative analysis of the review texts rules out the interpretation that reviewers focused on potential impact but failed to predict it. Instead, reviewers focused on the soundness of the manuscripts' arguments. We discuss how organizational characteristics of review can decouple reviewers' judgments from impact.

Introduction

Peer review is a crucial mechanism of resource allocation in science. Scientists and institutions depend on the judgments of peer reviewers to distribute billions of dollars and the limited pages of academic journals. It is thus important to understand how reviewers reach their decisions, and whether the decisions are reliable and valid. While the typical confidentiality of peer review has greatly constrained research, the literature has nonetheless grown to include hundreds of studies (Bornmann, 2011, reviews recent advances). The great majority of these studies focus on reliability – to what extent reviewers agree with one another (Bornmann, Mutz, & Daniel, 2010; Cicchetti, 1991a; Cole, Cole, & Simon, 1981). Validity of the decisions has

[*] Co-authored with Von Bakanic, Department of Sociology and Anthropology, College of Charleston. Published originally in the journal *Socius* (Teplitskiy & Bakanic, 2016). We thank the participants of the Social Theory and Evidence Workshop, University of Chicago, and the Skat25 Conference, Chicago, for helpful comments on previous versions of this research. We also thank the anonymous reviewers, Ben Merriman, and Clark McPhail for their input. All errors are our own.

proven a much more difficult target. Here we focus on the validity of manuscript review decisions in sociology.

Peer review is commonly conceptualized as the statistical problem of inferring the true value of a scientific claim from a noisy signal (Lee et al., 2013; Mervis, 2015). To assess the validity of decisions regarding a claim, one must then identify a claim's true value, or at least to define the characteristics of the evaluation process that produces unbiased value judgments (Lee et al., 2013). Yet in the social sciences and humanities, "value" may take on a variety of meanings. For instance, good work may be that which is competently conducted, or uses data creatively, or asks a novel question (Joshua Guetzkow et al., 2004; Michèle Lamont, 2009b). In interpretive disciplines scholars may define good work as that which impacts a conversation (Lamont, 2009: 61, 72). True value may thus be a multi-dimensional concept, and evaluating whether peer reviewers "get it right" necessitates specifying a dimension of interest, or at least how the various dimensions are to be made commensurable (Lee, 2015).

One dimension of value – a claim's potential for impact – is of special interest. First, it is often explicitly desired by funders and publishers of science. For example, Michael Lauer, a director with NIH's National Heart, Lung, and Blood Institute, "explicitly tell[s] scientists [that potential for impact] is one of the main criteria for review" (quoted in (Mervis, 2014). Second, impact is a unique dimension of value in that it is also a reward, and can thus promote a virtuous cycle. In settings where rewards are reputational, such as academia (Bourdieu, 1975), an excellent reputation requires visibility, which is largely shaped by academic journals (Clemens, Powell, McIlwaine, & Okamoto, 1995). Journals bring work to the attention of peers and, if peers cite the work, they make it yet more visible, incentivizing more such work in the future. It is thus desirable that the several dimensions of value correlate (are rewarded) with impact. Third,

impact is plausibly and easily measured by the classic, if beleaguered, metric – citations. Although scientists' motivations for citing the existing literature vary (Bornmann & Daniel, 2008a; Nicolaisen, 2007), citations are widely considered a measure of academic impact (Mingers & Leydesdorff, 2015; Raan, 1996).

A number of studies in the life and natural sciences have used citation impact to evaluate review decisions[1]. The logic is attractive: for many tasks, decisions of a crowd are more accurate than decisions of a few individuals, so perhaps citing decisions of a crowd of scientists represent a suitable gold standard against which to evaluate (the few) reviewers' decisions (Bornmann & Marx, 2014; Lee et al., 2013). The present study extends this line of research for the first time to sociology, where evaluation is among the least understood steps of the research process (Leahey 2008). We begin with the question:

Research question 1: Do reviews of published articles predict citations?

To answer this question we use the review files *American Sociological Review* (*ASR*), the flagship journal of American sociology, and focus on 167 research articles published between 1978 and 1982 (V. Bakanic et al., 1987). The age of the data allow us to evaluate how well peer reviews predicted citations over virtually the articles' entire lifetimes - 32 years following publication.

Although impact is an appealing gold standard of quality, in settings where value is multi-dimensional it is important to understand the criteria that motivate citers' and reviewers' decisions. In practice, analysts do not measure these criteria and assume them instead. For instance, it is often assumed that the audience of potential citers and reviewers decide whether to cite or to review positively, respectively, using identical criteria; In contrast, we focus on the outcome of citers' behavior, regardless of motivations, and measure reviewers' criteria. We ask:

[1] Prominent recent examples include (Li & Agha, 2015; Siler, Lee, & Bero, 2015).

Research question 2: What dimension of quality do reviewers focus on? Do they prioritize potential impact?

To answer this question we turn to the texts of the reviews, analyzed previously by Bakanic and colleagues (V. Bakanic et al., 1989). By assessing whether reviewers focus on impact or other dimensions of value we are better able to interpret the relationship between reviews and citations. In particular, if reviewers fail to predict impact, we can adjudicate whether they prioritize impact but fail to predict it fail, or if they focus on dimensions of value that are uncorrelated with impact[2].

The rest of this paper is structured as follows. First, we review existing studies that use citations to evaluate peer review validity. Second, we describe our data and method. Third, we present our results and discuss the possible interpretations. We finish with a discussion of policy implications.

Peer review, validity, and citations

Reliability and validity

Despite the common metaphor of peer review as a "black box," hundreds of studies have followed Robert Merton and colleagues' pioneering effort to investigate peer review processes[3]. The majority of this literature focuses on reliability -- how much reviewers of a particular manuscript or grant application agree with each other (Bornmann et al., 2010; Cicchetti, 1991a). Studies of the validity of peer review are rarer, and nearly absent in social science. For instance, of the 18 studies of grant review identified in (van den Besselaar & Sandström, 2015), nearly all focus on the life and natural sciences. The rarity is unsurprising: evaluations of validity require

[2] Reviewers may also fail to accurately assess these other dimensions. Our data does not allow us to address this possibility.

[3] (Zuckerman & Merton, 1971). For recent reviews see Bornmann, 2011; Lamont, 2009; Lee, Sugimoto, Zhang, & Cronin, 2013.

not only the reviewers' decisions but also a metric against which these decisions can be compared.

Accepted vs. rejected

Using citations as the gold standard of article quality requires trade-offs. In the case of academic journals, citations are usually observed only for the manuscripts judged publication-worthy. Negatively judged manuscripts either disappear from purview altogether, or reappear as articles published in other (usually lower impact) journals (Bornmann, 2011: 219). If initially-rejected articles accrue fewer citations after being published elsewhere, is the cause their lower quality or the publishing journal's lower status? Or, if citations *increase* after publication elsewhere, perhaps the articles were revised in a way that increased citability[4]?

In his review of the scientific peer review literature Lutz Bornmann found only 5 studies that compared citations received by manuscripts published and rejected by particular journals and 6 studies comparing citations to papers from funded and rejected grant proposals (2011: 218-25). These 5 journal-based studies all found that rejected manuscripts receive fewer citations than accepted ones (Bornmann 2011: 222). However, none[5] of these studies took into account that rejected manuscripts, if ultimately published, are generally published in lower impact journals. The findings from the grant and fellowship competitions were contradictory, which Bornmann speculates is due to the more prospective nature of these competitions (230). (Calcagno et al., 2012) surveyed a population of authors regarding submission patters and

[4] Research on revisions is limited and the impact of revisions on citation is unknown. Goodman and colleagues find that revisions reduce "spin" (Goodman, Berlin, Fletcher, & Fletcher, 1994). Decrease in spin may be expected to *decrease* citations. Two studies of sociology journals find that revisions alter the framing of results (Strang & Siler, 2015; Teplitskiy, 2015a), but the implications of change of frame on citations are unclear.

[5] Bornmann & Daniel (2008) come closest to properly taking publishing journals' impact factors into account by comparing published and rejected-and-published articles' citations to a baseline of *all* journals within the discipline of chemistry or one of its subfields. Unfortunately this approach does not preclude the possibility that journals publishing initially rejected manuscripts are still of higher impact factor than the field's average.

obtained data on 80,748 articles published in 923 biomedical journals. They found that rejected articles ultimately published elsewhere received significantly *more* citations than articles accepted by the first journal. It is unclear if these results are biased by a low response rate (37%). In another study, (Siler et al., 2015) found that articles *desk-rejected* at top biomedical journals and published elsewhere received fewer citations than articles rejected *after review* and published elsewhere[6].

Accepted vs. accepted

The approach taken here focuses only on the accepted manuscripts. This approach avoids the problem of variable journal prominence but suffers from limited variation in review scores: only high-scoring items are published or awarded. Considering only published articles can nevertheless yield a valuable and valid interpretation *for articles exceeding a minimum threshold of judged quality*. The generalizability of conclusions to the review process *as a whole* naturally depends on the variation in review scores *observed*: the higher the variation in quality among accepted manuscripts, the better the results should generalize to all reviewed manuscripts.

The evidence from "accepted vs. accepted" studies has been mixed. Siler and colleagues (Siler et al., 2015) used data from elite medical journals and found no statistically significant association between review scores and citations of published articles. (Danthi et al., 2014) examined the relationship between percentile rankings of 1491 National Heart Lung and Blood Institute (NIH) grants and 16,793 publications resulting from these grants. The authors found no relationship between grant percentile and citation metrics; however, a machine learning algorithm *was* able to identify citation "hits" from review scores.

[6] It does not appear that the authors accounted for the possibility that desk rejected articles were resubmitted to lower-impact journals than articles rejected after review.

In perhaps the largest study of its kind, (Li & Agha, 2015)) found that among publications generated from the 137,215 NIH grants a 1-standard deviation increase in percentile ranking was associated with 17% more citations. In another study of NIH, (Park, Lee, & Kim, 2015) used as a natural experiment the unexpected federal stimulus funds in 2009 that enabled the NIH to fund some proposals that had been initially rejected. They found that initially-rejected proposals produced 6.8% fewer citations per month than initially funded (higher rated) proposals. In sum, results from studies of peer review validity are mixed. In most cases initially rejected articles receive fewer citations than accepted ones, but the impact of the publishing journal on citations is generally not taken into account. In some cases there is not a relationship between citations and the review scores of accepted articles or grants, but in two important studies there is a statistically significant, albeit small, relationship.

By and large, the aforementioned studies have assumed away the motivations of reviewers and citers and interpreted the association between review scores and citations as an unproblematic referendum on the effectiveness of peer review as a whole. In contrast, we take into account the criteria used by the reviewers and develop additional interpretations. We argue that the extent to which reviewers are representative of the audience and what the reviewers take as their mandate can both mediate the relationship between reviews and citations.

Data and methods

Between 1971 and 1981, 2,337 manuscripts were submitted to the *American Sociological Review* (*ASR*)[7]. A sample of these manuscripts and their review data were originally used to evaluate several aspects of *ASR*'s review process, including fairness and reliability (V. Bakanic

[7] 5% of incoming manuscripts were desk rejected or withdrawn by the author(s). Of the remaining manuscripts 63% were rejected after review, 25% were revise-and-resubmit (and 19% eventually published), and 9% were unconditionally accepted (Bakanic, McPhail, & Simon, 1987: 634).

et al., 1987, 1989; Simon et al., 1986). Details of the data collection and coding process may be found in (Bakanic et al., 1987: 634-5). Here we focus on the 167 published full-length articles[8]. Additionally, Bakanic et al. (1989) analyzed the content of the reviews of 323 published and rejected manuscripts. We return to these content analyses here to assess to what extent reviewers focus on the potential for impact.

Reviewer selection and instructions

To select reviewers, the editorial staff considered three primary criteria. The staff first looked for reviewers whose work was in the same substantive area and, upon identifying suitable individuals, took into account the quality of their previous reviews (if any) and turn-around time. They excluded people who were known co-authors with, or students of the author(s). They avoided (if possible) sending a review to a colleague in the same department and/or institution (McPhail, 2016).

Reviewers were mailed a cover letter from the editor's office, a copy of the manuscript and a review form with instructions and a list of possible recommendation (reject, revise and resubmit, accept conditional and accept). The instructions requested reviewers to comment on the appropriateness of the manuscript for the *ASR*, as well as the quality of the literature review, theory, methods, analysis, conclusions, organization, writing and any other aspect of the paper they though appropriate. The editorial staff met weekly to consider the manuscripts with completed reviews. They also discussed manuscripts that for which they had trouble finding reviewers (McPhail, 2016).

[8] This set includes nearly all of the articles published between 1978 and 1981, and some published in 1982. We exclude from analysis research notes, comments, and replies. Citation information could not be located for 4 articles. For 45 articles information about the rounds of review was available but reviewers' recommendations were missing.

Review outcomes: consensus and average review score

Reviewers' recommendations in a particular round were dichotomized as the *consensus* variable, taking on the value "consensus accept" when all reviewers chose "accept" or "conditional accept" on their referee forms and "no consensus" when at least one reviewer choosing "revise-and-resubmit" or "reject". In one case the editor overrode reviewers' unanimous decision in the last round of "revise-and-resubmit." We also converted reviewers' recommendations into numerical scores using the scale 1=reject or refer, 2=revise and resubmit, 3=conditional accept, 4=accept. Each article's *average review score* is the mean (numerical) recommendations made by reviewers in the last round of review.

First vs. last round

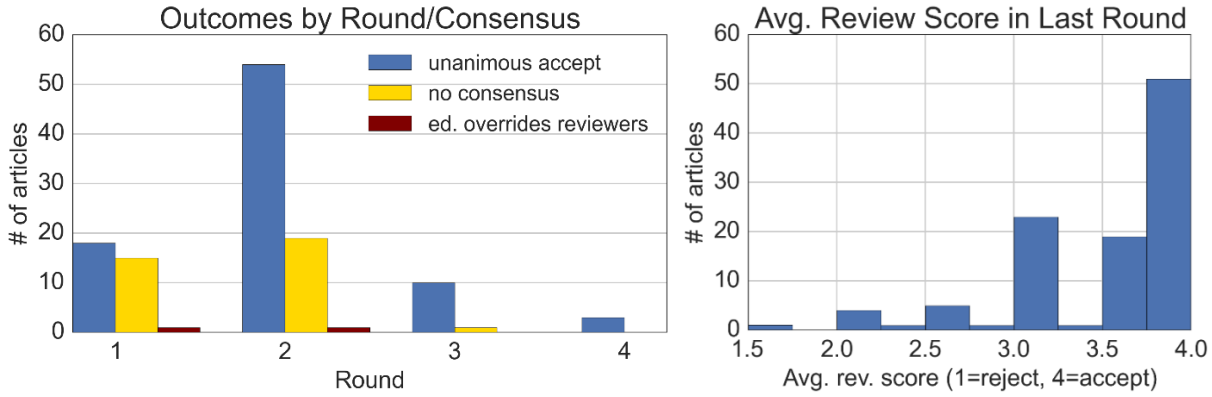
Most manuscripts underwent two or more rounds of review, with review scores assigned in each round[9]. Consequently, the analyst faces a choice between using scores assigned in the initial round of review or those given in a later round. Authors revise manuscripts between rounds, and it is not unusual for the changes to be substantial (Teplitskiy, 2015a). Reviewers' opinions of the final manuscript may have little to do with their opinions of its initial form, and it is the final version which accrues citations. Consequently we chose to focus on reviewers' scores from the last round (an analysis of citations vs. *first round* review scores may be found in the Appendix, Figure A3.1). On the other hand, scores given by a particular reviewer across rounds are generally constrained to increase. A reviewer who realized that her initial score was mistakenly favorable would find it very awkward indeed to explain to the editor and authors why, after the authors undertook revisions, a score worsened (and why the reviewer was not

[9] *ASR* would generally send a revise-and-resubmit manuscript to the original reviewers and one new reviewer. See (E. Y. Bakanic, 1986) for details.

thorough enough the first time). In sum, last round's scores have the merit of a tighter connection with the final article and the demerit of smaller variation (biased upward) in scores.

Figure 3.1 below summarizes the distributions of review outcomes.

Figure 3.1. Review outcomes of published articles. Left: Distribution of review round and consensus among reviewers. Right: distribution of average review score in the last round.



.

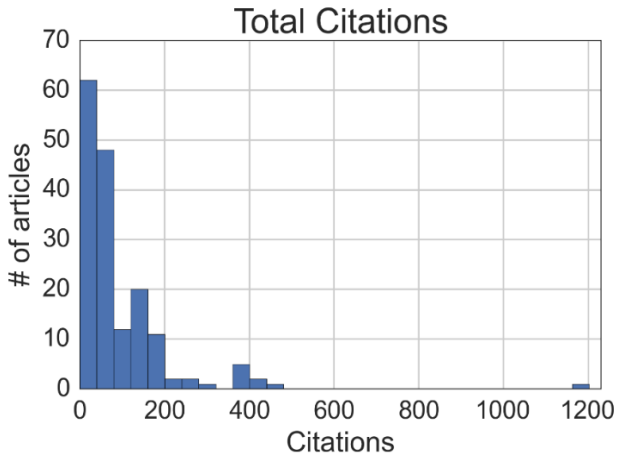
In the left panel of Figure 3.1 is the number of (published) articles in each round and consensus type. In the right panel of the figure is distribution of (published articles') *average review score*.

Total citations

Citation received by each manuscript in each year after publication were obtained from ThompsonReuters' *WebOfScience* [10] database between 2014-05-23 and 2015-02-17. Total citation counts include citations received by each article within the first 32 years after publication. Figure 3.2 below displays a histogram of total citations.

[10] Because the *WebOfScience* has poor coverage of the social sciences relative to other prominent databases, it may fail to index manuscripts *citing* the *ASR* articles under consideration here (Bornmann, Thor, Marx, & Schier, Forthcoming). This consideration should not affect the comparison of *WebOfScience* citation counts to each other, as all counts are disadvantaged equally.

Figure 3.2 Total citations for published manuscripts 32 years after publication.



The histograms of total citations counts are greatly left-skewed; the obvious outlier with 1204 citations is Cohen and Felson's (Cohen & Felson, 1979) article "Social Change and Crime Rate Trends: A Routine Activity Approach." Tables A3.1 and A3.2 in the appendix list the 10 most- and least-cited articles in the data sample. We address the skew of citations in the following analyses by log-transforming counts or using non-parametric measures (i.e. ranks) and statistical tests (i.e. Kolmogorov-Smirnov).

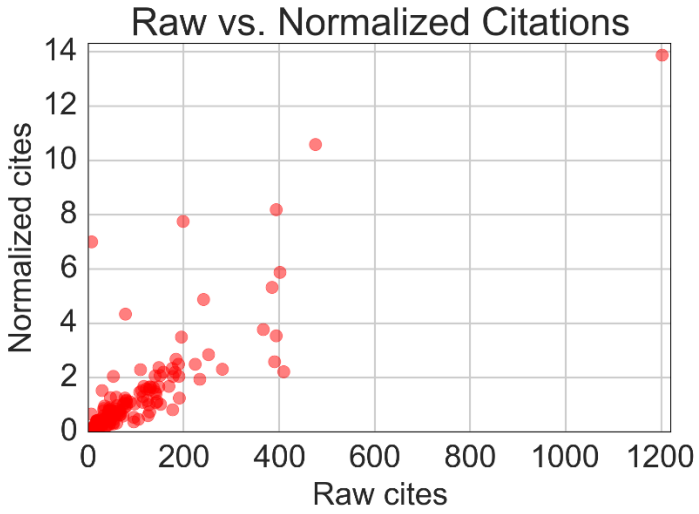
Subfield-normalized citations

Research fields vary in how frequently typical items are cited. It is thus one of the "key principles of citation analysis that citation counts from different fields should not be directly compared with each other" (Waltman, 2015). Sociology is a notoriously heterogeneous discipline (Smelser, 2015), and *ASR* welcomes and publishes material from all of its subfields. To compare citations across these various subfields we normalize citations by subfield in two ways. First, following common the bibliometric practice (Waltman, 2015), we define for each article a reference set that consists of other articles on the same substantive topic. The topics

consisted of the 54 sections of the American Sociological Association and 9 additional topics[11]. Coders classified each manuscript as belonging to up to three of these topics, e.g. “medical” or “theory”. All manuscripts sharing at least one of these subfields were included in the reference set.

Normalized citations carry information not present in the raw citations -- the two quantities correlate weakly (0.13), as is apparent in Figure 3.3, which displays their scatterplot.

Figure 3.3. Scatter plot of raw versus normalized citations 32 years after publication. Each dot represents an article. Normalized citations were obtained by dividing raw citations by the mean citations of articles in the reference groups – all those substantively oriented to the same ASA section. The correlation between the two quantities is 0.13.



Results

This section examines the relationship between reviews and (un-normalized) citations first, and then uses normalized citations. Next, two analyses probe whether especially high citations are associated with the highest scores. Finally, review texts are used to argue that

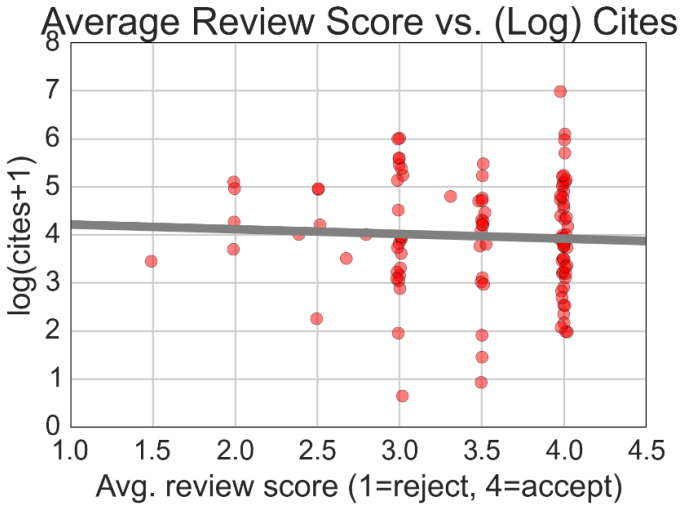
[11] For the complete list of topics and coding instructions, see (E. Y. Bakanic, 1986).

reviewers focus on the soundness of manuscripts' arguments rather than their potential for impact.

Review scores and citations

Figure 3.4 presents a scatter plot of *citations* vs. *average review score* (in the last round) with a line of best fit. (A similar scatter plot with review scores from the *first* round of review may be found in the Appendix, Figure A3.1).

Figure 3.4. A scatter plot of citations vs. average review score for each manuscript (red dot) and a line of best fit (gray). The correlation is = -0.051. The slope is -0.098 and statistically insignificant at the 0.05 level. The dots are jittered to improve visibility.



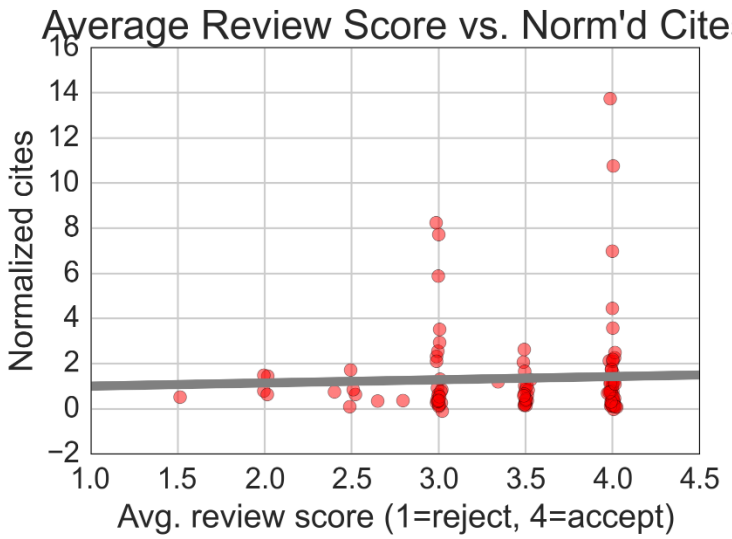
No relationship is apparent visually or statistically between the reviews and citations: the slope of the regression line is -0.098 (s.e. 0.187, p = 0.601).

Review scores and normalized citations

Does the apparent lack of a relationship between reviews and citations reveal a null relationship, or is a small but significant true relationship swamped by other determinants of citations, particularly target audience size? After all, an outstanding article (high review scores) targeted at a relatively small community may receive fewer citations than a mediocre article (low

review scores) targeted at a relatively large community. Unmeasured variation in audience sizes may introduce a large amount of noise in citation counts and make all but the strongest relationships appear null. To take into account audience size, citation counts were subfield-normalized by dividing citations by the mean citations received by other articles on the same substantive topic. Figure 3.5 displays a scatter plot and line of best fit for these (total) normalized citations vs. review scores.

Figure 3.5. A scatter plot of normalized citations vs. average review score for each manuscript (red dot) and a line of best fit (gray). The correlation is = 0.040. Citations were normalized by dividing each article's citations by the mean citations received by all other articles on the same substantive topic. The slope is =0.14 and statistically insignificant at the 0.05 level. The dots are jittered to improve visibility.



The relationship in Figure 3.5 echoes that of the previous figure: no relationship is apparent between review scores and normalized citation counts.

Multi-variate analysis

Bi-variate analyses displayed in Figures 3.4 and 3.5 show no evidence of a relationship between review scores and citations. However, the bivariate relationships did not control for important predictors of citations identified or posited by the literature. These predictors include

the number of co-authors (Wuchty, Jones, & Uzzi, 2007), institutional prestige (Leimu & Koricheva, 2005; J. S. Long, 1978), and subfield-spanning (Leahey & Moody, 2014). To control for these covariates, we estimated the following regression[12]:

$$NORM.CITES = \beta_0 + \beta_1 REV.SCORE + \beta_2 NUMB.AUTH + \beta_3 AUTH.PRESTIGE + \beta_4 AUTH.RANK + \beta_5 SUBFIELDS + \varepsilon$$

NORM.CITES (normalized citations) and REV.SCORE (average review score) were defined earlier. AUTH.PRESTIGE (author prestige) was measured as the prestige of the first author's current institution, 1=high school, 5=MA-granting institution, 10=PhD-granting institution with an ACE ranking of 60 or higher. AUTH.RANK (author rank) was measured as the first author's professional rank, with 1=undergraduate, 5=research associate, 9=professor emeritus. SUBFIELDS was measured as the number of substantive research areas the manuscript addressed (see footnote 11).

Table 3.1 displays estimates of this regression.

[12] Pair-wise correlation coefficients between the variables are reported in the Appendix Table A3.3.

Table 3.1. Coefficient estimates from the regression of normalized citations on review score and several author and manuscript characteristics identified in the existing literature as especially pertinent.

Dependent variable: **NORM.CITES** (normalized citations)		
Variable	Coefficient	95%-confidence interval
REV.SCORE	-0.23	-2.56, 3.07
NUMB.AUTH	0.74***	0.20, 1.28
AUTH.PRESTIGE	-0.030*	-0.065, 0.004
AUTH.RANK	0.038*	0.002, 0.073
SPANNING	-0.49	-1.18, 0.19
Number of complete cases = 106, $R^2 = 0.16$, Asterisks designate the following p-value thresholds: *=0.1 **=0.05, ***=0.01		

The central quantity of interest – the regression coefficient of the average review score – is statistically insignificant. Meanwhile, the number of co-authors is the strongest and most reliable predictor of citations. This multi-variate analysis provides further support to the earlier bi-variate analyses: even when controlling for important covariates, the data do not provide evidence of a relationship between review scores and citations.

Do peer reviewers identify citation "hits?"

If reviewers do not predict the impact of accepted articles overall, do they successfully identify citation "hits" –articles that accrue an uncommonly large number of citations (Danthi et al., 2014)? If reviewers and editors judge a submitted manuscript to be of uncommonly high potential for impact, they will likely wish to make its path to publication as quick as possible, and accept it after the first round of review. In our data there are 18 such "first round consensus accept" articles. Comparing citations received by this group – the "cream of the crop" – to

citations received by the 149 articles taking other paths to publication would then indicate whether unusually favorable review outcomes correspond to unusually high citations.

The distribution of citations is skewed, as noted previously. To assure that results are not unduly influenced by one or two outliers, we convert articles' raw citation counts into ranks: the article receiving the least citations is ranked 1 and the article receiving the most is ranked 167. It is then possible to observe whether one of the two article groups – 1st round consensus accept articles vs. all other articles – is overrepresented in the low or high article ranks. Figure 3.6 displays these densities using kernel density estimates (KDE) plots[13].

Figure 3.6. Density of articles across citation ranks by peer review outcome. Articles are split into two groups based on review outcome – 1st round consensus accepts vs. all others – and the presence of each group across the range of ranked citations is plotted. The densities shown are kernel density estimates.

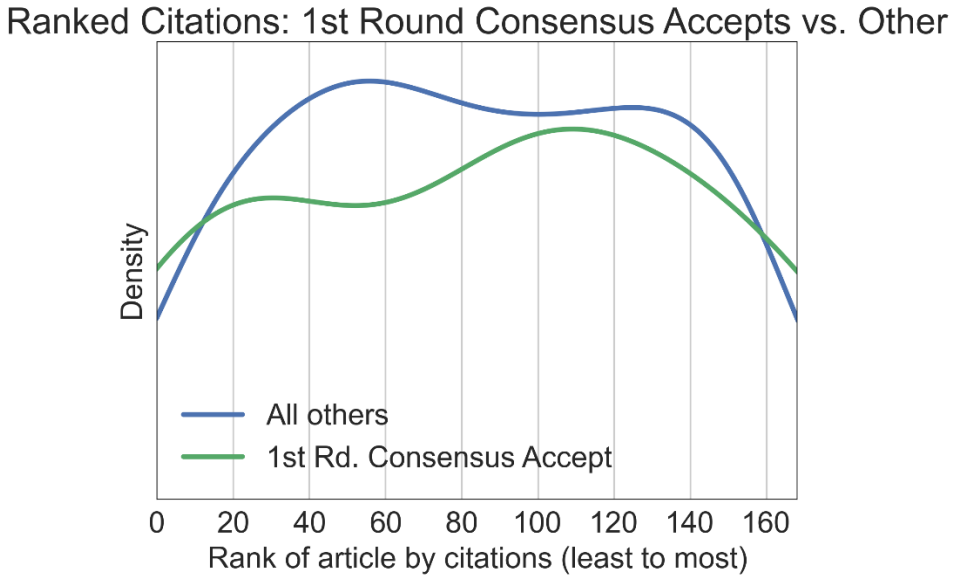


Figure 3.6 demonstrates that that 1st round consensus accept articles are not overwhelmingly represented among the high article ranks. In fact, the small overrepresentation among the most highly cited articles is balanced by a small overrepresentation among the least cited articles.

[13] KDE plots, like histograms, visualize distributions. https://en.wikipedia.org/wiki/Kernel_density_estimation. Accessed 2015-10-22.

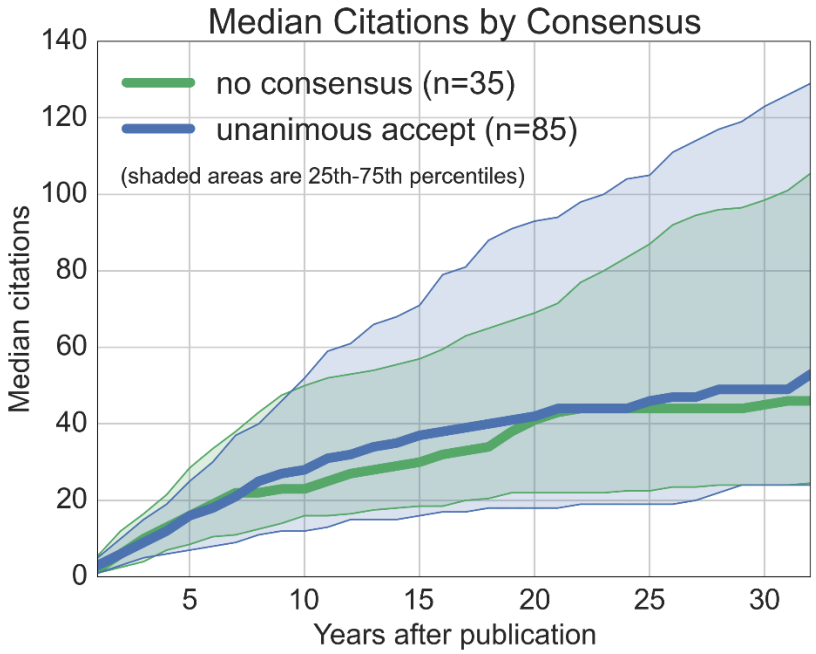
A statistical test confirms the visual evidence. The Kolmogorov-Smirnov (KS) test is a nonparametric (unaffected by skew) test of whether two probability distributions differ in any part of the distribution. A KS test of whether the distribution of (1) 1st round consensus accept articles' citations differs anywhere along the distribution from (2) all other articles' citations does not reject the null hypothesis of no difference (*KS statistic* = 0.17, *p-value* = 0.73)[14]. We thus find no evidence that review "hits" are associated with citation "hits."

Impact at each year after publication

The age of the data invites an analysis of not only whether review scores predict accepted manuscripts' *total* citations, but also citations at each year after publication. The manuscript were divided into two groups: *consensus accept* manuscripts are those which all reviewers recommended for acceptance, while *no consensus* manuscripts are those which at least one reviewer recommended be rejected or revised-and-resubmitted. Citations were obtained for each year after publication. Medians were used instead of means to limit the influence of one outlier (using means does not qualitatively affect the conclusions). Figure 3.7 displays the median citation trajectories of these *consensus accept* and *no consensus* groups.

[14] We also estimated logistic for the relationship between average review score and whether an article's citations fell (=1) or did not fall (=0) in the top 5% or 10%. In both regressions the average review score was not associated with the odds of the article representing a citation "hit." Results are available upon request.

Figure 3.7. Median citation trajectories by consensus. Solid lines: cumulative median citation trajectories of consensus accept and no consensus articles. Review decisions are from the last round of review. Shaded areas: 25th – 75th percentile trajectories of each group.



The visual evidence is clear: the two groups of articles, one with superior reviews and one with inferior reviews, have nearly indistinguishable citation trajectories over the entire 32 year period[15]. The variance of both groups dominates the minor (and unstable) differences in their medians. Thus, the apparent independence of reviews and citations 32 years after publication established earlier holds throughout the observation period.

Do reviewers value citation impact but fail to predict it?

All of the results above point unanimously to the conclusion that review outcomes of accepted articles do not predict their citation impact. The lack of a relationship raises a question: Do reviewers *value* citation impact but fail to predict it or do they privilege other manuscript characteristics, such as soundness of argument? The review instructions offer little guidance – as

[15] Statistical tests comparing the mean citations (or mean logged citations) of the two groups at each year lead to the same conclusion and are available from the authors upon request.

discussed in the Data and Methods section, reviewers were asked to comment on the "appropriateness" and "quality" of the manuscript. These rather open-ended reviewer instructions are not unusual for sociology. Furthermore, sociologists' formal training rarely includes review instruction, making reviews practices difficult to deduce from the *instructions* and potentially non-uniform. As one sociologist put it, "Manuscript reviewing must be one of the most important, least formally trained professional functions that we serve" (Brunsma, Prasad, & Zuckerman, 2013).

The focus of reviewers' attention has usually been elicited with surveys and interviews (e.g. Bornmann, Nast, & Daniel, 2008; Brunsma et al., 2013; Chase, 1970; Lamont, 2009). A chief problem with eliciting-by-asking, in addition to desirability effects, is that the reviewers are not forced to compromise between criteria in the way they must when actually reviewing. There is no realistic constraint forcing respondents to trade-off among criteria: they may simply respond that they value them all. These considerations motivate our choice to deduce reviewers' foci of attention from the actual review texts. But before we turn to these texts, consider what reviews would focus on if citations truly were of utmost importance.

Expert reviewers are likely to have in mind mechanisms that generate citations in their field. Indeed, impact in the field is to some extent a prerequisite of being the type of reviewer that *ASR* would call upon. Some reviewers may even be familiar with the voluminous scholarly literature on the predictors of impact. If mandated to predict citations, reviewers would likely utilize all this knowledge in evaluating a manuscript. They would hesitate to accept a well-executed piece of scholarship if it was unlikely to generate citations, perhaps because an earlier article solidified the first-mover citation advantage (Newman, 2009) or because it did not address several audiences (Leahey & Moody, 2014). A manuscript from poorly positioned authors or

research areas could also drag citations down (Bornmann, Schier, Marx, & Daniel, 2012) and would thus be a legitimate cause for rejection. Articles presenting useful methods, which are overrepresented among the most cited articles, would win praise easily (Van Noorden, Maher, & Nuzzo, 2014). And of course reviewers could write directly, "I am afraid this manuscript will not be cited."

The contents of the *ASR* review reports did not mention citations explicitly and even implicit mentions were exceedingly rare. The vast majority of review text in these data, whether comments to the author or editor, whether praise or criticism, is devoted to general impressions and to concerns about the soundness of a manuscript's argument (V. Bakanic et al., 1989).

Positive comments were relatively rare and, when made, were most often generic (36.8% of all positive comments) (Bakanic et al. 1989: Figure 2). For example, reviewers praised manuscripts by calling them "interesting" (pg. 644) or writing "On the whole this is a nice article" (pg. 645). Reviewers did often (19.9%) praise the manuscript's topic. For example, one reviewer wrote, "This is an important topic for studies of subjective class identification" (pg. 645). Such comments on topic may be *indirect* endorsements of the manuscript's citability[16]. In sum, even if style (9.7%) and topic (19.9%) are taken to be endorsements of citations, nearly 70% of reviewers' positive comments focused on either very general impressions or the argument's soundness.

Reviewers' negative comments to the *editor* were most often general (31% of all criticisms). These included statements from "This is not sociology" to "I wasn't convinced by this research" to, even, "Yuk!" (pg. 646). Reviewers provided to the *authors* more concrete

[16] For example, Bornmann and colleagues (2012) studied citations of manuscripts published by a chemistry journal and found that citations, when controlling for peer review score, were predicted by the chemical subfield.

criticisms, focusing on problems of theory (13.2%), analysis (12.3%), and results (11.7%) (pg. 647). Criticisms of the manuscript's importance to the field, which may be construed as related to its citability, are relatively rare: 6.4% of comments to the author and 6% of comments to the editor criticized the manuscript's topic. Similarly, criticisms of style, perhaps also related to citations, constitute 10.8% and 11.9% of comments to the author and editor, respectively.

The texts of reviewers' reports thus indicate that reviewers do not focus on impact. They do not mention citations or other measures of impact *explicitly* at all – and if impact was understood to be part of their mandate, how could they not? Furthermore, comments that may implicitly relate to citations, i.e. regarding a manuscript's topic, are much rarer than comments regarding the soundness of argument. Instead, reviewers take as their chief mandate to evaluate the soundness of a manuscript's argument, addressing most comments to matters of theory, measurement and analysis.

In summary, *ASR* reviewers did not make invalid predictions of impact. Instead, reviewers assessed the soundness of manuscript and, for accepted manuscripts, these judgments are unrelated to citations.

Discussion

This study explored the relationship between accepted manuscripts' peer review scores and citation impact. In addition to a possible linear relationship between scores and raw citations, the study examined citations normalized by subfield, citation "hits", and citations at each year in the 32-year span after publication. In none of these specifications did the data manifest a relationship. Did reviewers value potential impact but fail to predict it? Nearly all previous studies did not measuring reviewers' motivations directly and interpreted reviewers' ability or

inability to predict citations as a referendum on the validity of peer review (Bornmann, 2011b; Mervis, 2015). In contrast, this study used the texts of the reviewers' reports to deduce the foci of their attention. The texts show that reviewers did not prioritize potential impact directly or indirectly; instead they focused on the soundness of arguments. Reviewers' judgments of soundness, not potential impact, are unrelated to citations.

Although this study helps rule out one interpretation of the absence of a relationship between scores and citations, several interpretations remain. First and foremost, the true relationship between review scores and citations may be nonlinear – there may be a threshold of quality below which a strong relationship exists but above which it disappears. This may be the case where the chief distinction (threshold) in quality lays between rejected and accepted (whether immediately or after revisions) manuscripts, while gradations in quality among the accepted articles are inconsequential (Mervis, 2015). Two additional interpretations have been largely overlooked by the literature and are discussed below.

Non-random reviewer selection

Reviewers are not selected randomly from a population of potential citers, so their judgments may fail to generalize to this population. Both reviewer solicitation and agreement-to-review predispose the final reviewer panel to consist of individuals who find the manuscript's research area valuable. First, editors at *ASR* and likely in other settings solicit reviews from experts in the field from which the manuscript comes. These experts would not have made the enormous investments required to master their specialties if they did not find the specialties valuable. Second, research on motivations of volunteers in crowd-sourcing settings emphasizes that volunteers donate time in order to learn and be creative (K. R. Lakhani & von Hippel, 2003; K. Lakhani & Wolf, 2003). We suspect that in scientific peer review these motivations are

equally crucial: experts agree to perform the time-consuming and uncompensated task based, in part, on interest in the research area. In sum, reviews come from individuals uncommonly interested in the manuscript's research area. Even superlative assessments from reviewers may thus fail to translate into citations from the broader population if it has no interest in the manuscript's topic.

Implicit division of labor between reviewers and editors

Conspicuously absent in our discussion and others in the literature are the editors. Editors may play a crucial role in mediating the relationship between reviews and citations not only because they select reviewers, but because the reviewers may rely on an implicit division of labor between themselves and the editors. In particular, some reviewers may take it as their mandate to evaluate manuscripts for soundness and leave judgments of whether it is interesting, judgments that may be strongly related to citations, to the editors. This implicit division of labor is apparent when one experienced sociological reviewer, in a survey of ASA reviewers, writes "…a fourth test [of the manuscript under review] is whether the point is interesting or innovative, but I feel that this can be left for the editor to decide so I don't make a big deal of it" (Brunsma et al., 2013). However, in the absence of formal reviewer training or explicit instructions, it is not surprising that reviewers vary in what they take to be their task. For example, in the same ASA survey another reviewer writes, "I simply read [the manuscript] from the perspective of a potential reader. The standard and maybe only thing I look for is this: Is this paper interesting? That is my prime criterion." (pg. 9). In sum, reviewers differ in their evaluative criteria and in how they divide the evaluation task between themselves and the editors. To the extent that reviewers expect editors to judge a manuscript's interestingness, the absence of a relationship

between accepted manuscripts' reviewer scores and citations is unsurprising: the crucial relationship to evaluate is between citations and editors' decisions.

Conclusion

This study focused on the evaluation of sociological manuscripts – a crucial and understudied research practice (Leahey, 2008a). We returned to the decisions made by *American Sociological Review* reviewers and editors more than three decades ago to assess whether the decisions predicted the published manuscripts' citation impact over what is essentially their entire lifetimes. The chief empirical finding is a null one: reviews do not predict citations in a number of plausible specifications. In addition to extending the empirical base of the literature to sociology, the study developed a number of interpretations of the relationship especially relevant to social science, where the value of manuscript excellence can be multi-dimensional. By combining bibliometric and qualitative data the number of these interpretations of sociological peer review was reduced. In particular, ruled out is the interpretation that reviewers prioritize impact but fail to predict it. Yet several interpretations remain. The relationship between review scores and impact may be non-linear, with an association at lower levels of quality and no association above a publishability threshold. Additionally, reviews may fail to predict citation because of non-random reviewer selection or because an implicit division of labor assigns the difficult task of predicting citations to the editors.

The age of these data enabled us to trance the fate of manuscript through an unprecedented length of time – 32 years, virtually the articles' entire lifetimes. Yet the age may limit applicability to contemporary academic publishing. It is possible that the bibliometric conception of impact prevalent today would have been distasteful or foreign to reviewers 35 years ago. The standardization and quantification literature suggests that rating and ranking

systems do not simply quantify existing reality; they also induce evaluated and evaluating persons to behave in ways that fall in line with the system of valuation undergirding those systems (Espeland & Sauder, 2007; Espeland & Stevens, 2008). The increased prevalence and ease of using bibliometrics to measure impact may have changed scholarly norms for assessing the value of work. It is thus plausible that reviewers today, more conditioned to think and evaluate in terms of bibliometrics, may be more successful forecasters.

Despite the limitations above, this study contributes to our understanding of peer review and policies that may improve it. One direction for improvement concerns reviewer instructions. If a journal prioritizes citation impact, reviewers should be explicitly instructed to evaluate on this dimension. Another direction concerns reviewer selection. In the case of the *American Sociological Review,* the two types of a manuscript's "evaluation," that by expert reviewers and that by a more general citing audience, appear decoupled. It is plausible that more representative review panels will result in higher associations between review outcomes and citations. To the extent that citation impact is valued editors may need to trade domain expertise for representativeness.

Appendix 3

Table A3.1. The 10 most cited articles in the data sample.

Author(s)	**Year**	**Title**	**Citations**
Lawrence E. Cohen and Marcus Felson	1979	*Social Change and Crime Rate Trends…*	1202
Ronald L. Akers et al.	1979	*Social Learning and Deviant Behavior: A Specific Test…*	476
Delbert S Elliott	1980	*Reconciling Race and Class-Differences in Self-reported and Official…*	410

Table A3.1 continued			
Nan Lin, Walter M. Ensel and John C. Vaughn	1981	*Social Resources and Strength of Ties: Structu...*	402
Karen Cook	1978	*Power, Equity and Commitment in Exchange Networks*	394
David A. Snow, Louis A. Zurcher, Jr. and Sheldon Ekland-Olson	1980	*Social Networks and Social Movements: A Microstructural Approach...*	394
Thomas A. Heberlein and Robert Baumgartner	1978	*Factors Affecting Response Rates to Mailed Questionnaires…*	390
E. M. Beck, Patrick M. Horan and Charles M. Tolbert II	1978	*Stratification in a Dual Economy: A Sectoral Model…*	385
Paul D. Allison	1978	*Measures of Inequality*	367
Kenneth A. Bollen	1980	*Issues in the Comparative Measurement of Political Democracy*	281

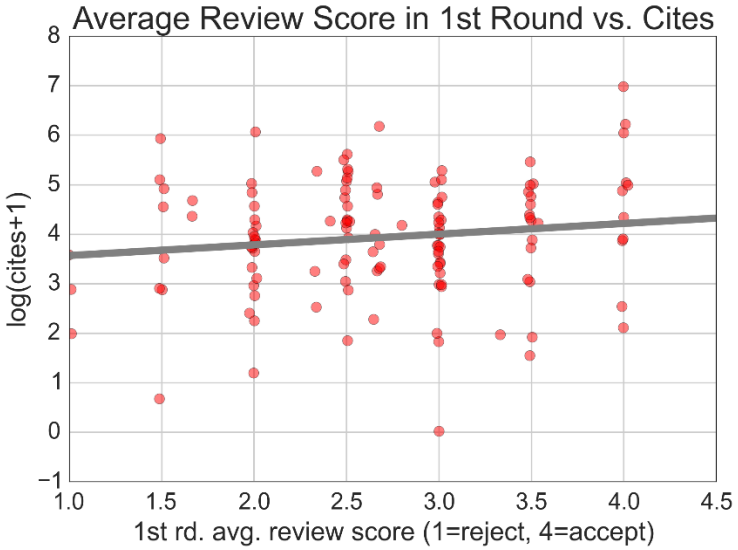
Table A3.2. The 10 least cited articles in the data sample.

Author(s)	**Year**	**Title**	**Citations**
Richard p. Applebaum	1978	*Marx's Theory of Falling Rate of Profit…*	0
Brian L. Pitcher	1978	*Diffusion of Collective Violence*	0
Ronnie Steinberg Ratner and Paul Burstein	1980	*Ideology, Specificity, and the Coding of Legal Documents…*	1
David E. Payne	1978	*Cross-National Diffusion: The Effects of Canadian TV …*	2
John W. Gartrell	1981	*Inequality within Rural Communities of India*	3
Israel Adler	1978	*Cross Pressures During Socialization for Medicine*	3
Colin Campbell	1982	*A Dubious Distinction? An Inquiry into the Value and Use of Merton's…*	4
G. Edward Stephan and Douglas R. McMullin	1981	*The Historical Distribution of County Seats in the United States…*	6
Michael C. Burrage and David Corry	1981	*At Sixes and Sevens: Occupational Status in the City of London from…*	6
William C. Rau	1980	*The Tacit Conventions of the Modernity School…*	6

Table A3.3. Pair-wise correlations between the variables used in the multi-variate regression.

	total citations	normalized total citations	average review score	number of authors	number of subfields	first author prestige	first author rank
total citations	1.000						
normalized total citations	0.128	1.000					
average review score	0.034	0.040	1.000				
number of authors	0.211	0.013	0.013	1.000			
number of subfields	-0.077	-0.202	0.052	-0.195	1.000		
first author prestige	-0.016	0.002	0.100	0.067	-0.013	1.000	
first author rank	0.003	0.007	-0.011	0.073	-0.090	0.489	1.000

Figure A3.1. Citations vs. review score in the initial round of review. The slope of the regression line equals 0.22 and is not statistically significant (s.e. = 0.14, p = 0.13). The dots have been slightly jittered to improve visibility.



CHAPTER 4. HOW FIRM IS SOCIOLOGICAL KNOWLEDGE? REANALYSIS OF GSS FINDINGS WITH ALTERNATIVE MODELS AND OUT-OF-SAMPLE DATA, 1972-2012*

Abstract

Published findings may be fragile because hypotheses were tailored to fit the data and knowledge about insignificant relationships—"negative knowledge"—remains unreported, or because the world has changed and once robust relationships no longer hold. We reanalyze findings from hundreds of articles that use the General Social Survey, 1972-2012, estimating (1) published models and alternative specifications on in-sample data, and (2) published models on future waves of the GSS. In both, number of significant coefficients, standardized coefficient sizes, and R^2 are significantly reduced. Our findings suggest that social scientists are engaged in only moderate data mining, but that they could benefit from more; a bigger concern is the relevance of older published knowledge to the contemporary world.

Introduction

Social scientists and policy makers rely on the social science literature to understand the social world, craft new research, and design policies to improve it. Yet many readers question to what extent this literature can be trusted. Some concerns stem from the proliferation of computing technologies that facilitate the unwitting use and abuse of statistical testing (King, Tomz, & Wittenberg, 2000; McCloskey & Ziliak, 1996; Morrison & Henkel, 2006). Larger storage and faster computation enables social scientific analyses to be performed in less time with less effort. This trend is of special concern given the long-standing publication bias towards

* Co-authored with James Evans.

"positive" or statistically significant findings (Rosenthal 1979; Franco, Malhotra, and Simonovits 2014; Gerber and Malhotra 2008).

With improvements in computation, social scientists who desire to produce statistically significant findings can often do so by estimating many model specifications until finding an idiosyncratic model that "works"—that yields a sufficiently low *p*-value. On random data, an analyst can "accidentally" produce a statistically significant relationship between an independent and the dependent variable at the $p<.05$ level by estimating an average of twenty models with different variables ($20 \times .05=1.0$). Many have argued that this practice, termed "asterisk-hunting," "*p*-hacking," or unreported "data mining"[1], is sufficiently widespread as to undermine the integrity of the social science literature (Freese 2007; Ioannidis 2005; Ioannidis and Doucouliagos 2013; Simmons, Nelson, and Simonsohn 2011). This trend also portends a growing pool of private insight about tested social scientific relationships that did not "pan out" and never achieve publication. Often called negative knowledge, insight about non-relationships is rarely stored, almost never shared, and so becomes at risk for repeated discovery or neglect. Every non-robust finding hides beneath it negative knowledge, the invisible "dark matter" of research.

Concerns about robust knowledge reach beyond the social sciences and have been voiced repeatedly across the academy, for example in neuroscience (Button et al., 2013), genetics (J. P. A. Ioannidis, Ntzani, Trikalinos, & Contopoulos-Ioannidis, 2001), machine learning (Pentland, 2012) and high energy physics (Lyons, 2013). A cover article from *The Economist* titled "Unreliable Research: Trouble at the lab" (*The Economist*, 2013) and several related popular

[1] While data-mining in the Computer Science community does not necessarily involve overfitting or "*p*-hacking", its usage in the social sciences implies not only searching for relationships, but reporting them without full disclosure of the "mining" process.

media features cite contemporary experts who argue that these and related practices have become commonplace in contemporary research, and have resulted in fragile findings and an unreliable published knowledge base.

Published sociological findings could also be fragile for another reason. Findings can weaken if the social world they once robustly described changes. Consider an example from *What's Wrong with Sociology*, in which Stephen Cole considered how social change stymies efforts to develop robust theory*:*

> *In 1970 only about 10 percent of all applicants to medical school were women and in 1983 one-third were women …. In 1970 an explanation of inequality could have referred to gender norms which proscribed high commitment careers such as medicine as being incompatible with the family roles of women. In 1983 these proscriptive norms had all but disappeared … In sociology by the time any theory is developed … it is possible that the phenomenon and the factors causing it will have changed.* (Cole 2001: 43).

This phenomenon of shifting realities may be as relevant as *p*-hacking to the robustness of the social science literature. How well findings hold up across time, however, has not been studied quantitatively and is not mentioned as frequently in discussions of reproducibility and robustness. Many other sciences encounter this same concern, like ecology and conservation biology (Nyssa, 2014), meteorology, climatology and other earth sciences where the world changes (James E. Lovelock, 1990; J. E. Lovelock, 1965), partially in response to public awareness of scholarship about it.

In this paper, we estimate the robustness of sociological findings in the published record and the extent to which it is affected by data-mining and a changing social world. We do this by computationally reanalyzing results from hundreds of articles that use the General Social Survey, 1972-2012. In order to test the relationship between robustness and data-mining, we compare published models to alternative model specifications (estimated on the same data) that swap the

original model with another that replaces one randomly selected variable with a semantically similar and highly correlated variable. This allows us to estimate the likelihood and degree to which scientists have selectively searched the space of models in the semantic neighborhood of those eventually published. Our investigation complements papers that reveal important and particular negative findings by estimating the overall prevalence of such findings hidden within published literature using the GSS. Unreported negative knowledge, however, is not the only reason for fragile sociological findings. Once-robust relationship can become weakened by shifts in the social world. In order to test the robustness of findings to social change, we re-estimate published models on all waves of the GSS collected after publication.

Our findings reveal that a moderate amount of unreported data-mining occurs in research using the GSS. Alternate model specifications estimated on the same data as published models do not fit the data as well: alternative specifications reduce the number of significant coefficients and model fit as captured by R^2 and adjusted R^2. When published models are estimated on future data collected after publication, the average size of coefficient and model fit significantly decrease, such that robustness "lost" to a decade of socio-cultural change is comparable to that which is "lost" to unreported data-mining.

The remainder of this paper is organized as follows. First, we describe the historical context through which findings in much of quantitative social science came to be about variables, regressions and effect sizes. Second, we review the literature on research practices and robustness. Third, we describe the details of our approach and findings. We conclude with a discussion of the significance these results hold for understanding the robustness of social science findings beyond the GSS.

Statistics and Social Science

A substantial branch of social science literature in recent decades attempts to estimate the strength of relationships or "effects" between variables representing social quantities. This form of social analysis and the view of the world it produces—what Andrew Abbott has called "general linear reality" (Abbott, 1988)—can be traced to the emergence of statistics in the 19th Century. When 19th Century policy makers and social scientists spoke of "statistics," they generally referred to compilations of official numbers, gathered through censuses, tax collection and the processing agencies of the state (e.g., number of prisoners; number of persons in asylums) (Camic & Xie, 1994). Later "index numbers" or special averages used to measure temporal fluctuations of wages, profits, and other important quantities were considered statistics (Christ, 1985; Persons, 1925; S. M. Stigler, 1978), but concern focused on trends in these numbers and not statistical inference.

The first major inferential approach to social statistics debuted with the Belgian astronomer Adolphe Quételet's 1835 *Essays on Social Physics*,[2] which translated work on error theory and sought to identify the "truth" by averaging across a range of observations to human data. Quételet devised the concept of the "average man" and used massive human measurements to reveal how mean values tend to follow a normal distribution[3] and that rates of behavior like crime and suicide stably persisted through periods of major upheaval (Beirne, 1987; Quételet, 1835). This average-based or "Continental approach" to statistics (Duncan, 1984) entered German experimental psychology (Gigerenzer, 1987; Porter, 1986; S. M. Stigler, 1986; Wundt,

[2] The full title, translated, was: *On Man and the Development of his Faculties, or Essays on Social Physics*. Auguste Compte coined "sociologie" to reject this approach to measurement, quipping that Quételet's measures were "simple statistics", which tarnished his own earlier use of "social physics" (Beirne, 1987; Comte, 1830).

[3] Two years later, Siméon Denis Poisson estimated the number of wrongful criminal convictions with what came to be known as the Poisson distribution, used to model any number of discrete occurrences within a given time-interval (Poisson, 1837).

1862, pp. 71–72) and just before the turn of the 20th Century was the first statistical approach promoted in American sociology at Columbia University in the writings of Franklin H. Giddings (Camic, 1995; Camic & Xie, 1994).[4]

The "science of variation" or "British approach" to statistical inference (Camic & Xie, 1994; Duncan, 1984) on which correlation and regression are based, emerged in the late 1870s with Francis Galton who redirected statistical thought from averages to the analysis of "variation for its own sake" (Porter, 1986, p. 129). He was not interested in the average man, but "men…different from the average"—those of superior and inferior human endowments and the hereditary basis of their differences (Hilts, 1973, p. 229). To Galton, exceptional human endowments were not "errors" and no less real than the average. Galton's interest in heritability led him to explore the relationship between the distributions of two or more factors, which led to the invention of correlation and regression in the 1880s (Hilts, 1981; Porter, 1986; S. M. Stigler, 1986, 1989). By the mid-1890s, general formulae, algorithms and extensions to Galton's statistical toolkit were added by Karl Pearson, George Udny Yule, and Francis Ysidro Edgeworth, including correlation coefficients and the least squares approach to multiple linear regression (S. M. Stigler, 1978, 1986).

Regression and correlation entered U.S. social science through anthropology and economics. Pioneering anthropologist Franz Boaz, at Clark and then Columbia Universities, emphasized the diversity of human types and the overlap of population traits rather than a focus

[4] Giddings argued that sociology's foundational concept was "consciousness of kind," which drove human self-organization into "categories . . . of real or supposed resemblance," "color, race, and nationality," "religious belief," and "political conviction" (1899, pp. 151–52). This category-centered view made categorical membership data central, and so appropriate sociological data were typically discrete or "absolute numbers" (1901, p. 29; 1910, p. 722; Camic & Xie, 1994).

on averages (Boas, 1894, p. 227, 1897, p. 151; Sokal, 1987; Stocking, 1968).[5] Henry L. Moore, also at Columbia, applied variational approaches to the study of wages, subsistence costs, living standards, market supply and prices using simple, partial and multiple correlation and regression (Camic & Xie, 1994; Moore, 1911, p. 19). Moore was first to "furnish empirical estimates of the parameters in theoretical models" by statistically controlling for confounding factors in natural settings (Moore, 1911, p. 23; G. J. Stigler, 1962; Camic & Xie, 1994).

Multiple correlation and regression were eventually acknowledged by Giddens and others at the beginning of the 20th Century as possible methods for sociology, and were gradually imported into sociology and specific fields of study, especially applied sociological research on crime and education. From the 1930s through the 1960s, Paul Lazarsfeld and what eventually became the Bureau for Social Research at Columbia University developed and promoted surveys, content and focus-group analysis. This made opinions, identities and behaviors (e.g., consumption, church attendance) available and suitable (Leahey, 2005) for measurement, statistical analysis and causal inference. Lazarsfeld also popularized these methods for business (e.g., marketing and media analysis) and social science, which received major government support following WWII and the rise of U.S. federal funding agencies in the 1950s (Jeřábek, 2001).

Shared computing technologies began to become available to social scientists in the 1960s and by mid-1970s sociologists performed statistical analyses on mainframe and desktop computers using software like SPSS and SAS (Leahey 2005: 6). The 1970s also saw the birth of one of the most important U.S. social surveys. Responding to the need for reliable, longitudinal

[5] Boaz pushed the envelope of statistical correlation (Stocking, 1968, p. 168) and precipitated analysis of variance techniques (Xie, 1988, p. 276).

social data, James Davis, then at Dartmouth, pushed development of national data collections on topics of broad social scientific interest. In 1972 this effort resulted in the General Social Survey (GSS). The GSS was a National Science Foundation-funded, nationally representative public opinion survey. This survey continues to be conducted by the National Opinion Research Center and has spawned similar surveys in a number of other countries. Davis' idea and effort has been an unequivocal bibliometric success: tens of thousands of research studies in sociology, political science, economics, and other fields have used these data to ask questions about both the changing "pulse of the nation" and fundamental social processes (Gibson, 2013). Because of its popularity and substantive breadth—ranging from religious attitudes and political opinions to educational aspirations—the GSS and the thousands of publications that have used it are an ideal setting in which to study negative and published knowledge.

Statistical Significance in Sociology

Social scientists and policy makers rely on authoritative and stable "effects" from the published social scientific literature to conduct research and design policy. Consider the "Equality of Educational Opportunity" Report (1966) requested by the Civil Rights Act of 1964, in which James Coleman, a student of Lazarsfeld, used the statistical analysis of student and school data to demonstrate that socioeconomic background overshadowed school funding as a predictor of students success. This deeply influenced the debate about segregation and the rise of desegregated busing systems.[6]

In recent years, concerns have grown over the robustness of findings resulting from statistical analyses of datasets like the GSS. The statistical significance of these findings is often

[6] In his 1975 analysis, Coleman published another evaluation that demonstrated how busing had failed in its desired effect by inducing "white flight" (Ravitch, 1978).

evaluated within the *p*-value paradigm, not only in sociology (Babbie, 2014), but economics, psychology (Wetzels et al., 2011), and biology. The *p*-value is used in Frequentist (and not Bayesian) inference: it is a function only of the observed sample results and not prior expectations. As such, it does not support explicit reasoning about the probabilities of hypotheses. Rather, the *p*-value provides a tool for testing a hypothesis by defining a test threshold that determines whether the analyst should accept the null or alternative hypothesis. The significance level or α is chosen, traditionally 5% or 1%, before performing the test (Leahey, 2005). If the *p*-value equals or falls below the significance level (α), the observed data is viewed as inconsistent with holding the null hypothesis true, and should be rejected in favor of the alternative. This guarantees that the Type I error rate or rate of false positives will be no greater than α if the *p*-value is calculated correctly.

The *p*-value is commonly interpreted as the probability of obtaining the observed sample results or some "more extreme" result, if the null hypothesis is assumed to be true (Hubbard, 2004). In sociology, a relationship between two variables is considered "statistically significant" if the probability that the apparent relationship was produced completely by chance is under a predefined threshold, usually 5% (Leahey, 2005). The use of statistical significance testing and the .05 *p*-value was first deployed in R.A. Fisher's book, *Design of Experiments* (1935) (Schmidt and Hunter 1997). In a recent article, which examined a twenty percent sample of publications in the *American Journal of Sociology* and *American Sociological Review*, 1935 until 2000 (1,215 articles), Leahey found of those that tested hypotheses using empirical, numeric data (613 articles) and statistical significance tests (496 articles), 86% used the .05 alpha level, 67% used .01, and 52% used .001 (Leahey, 2005). Ninety-six percent of those studies that could have reported statistical significance levels did report them, suggesting that this has become both a

social norm and a likely basis for evaluation. Because even plausible models often do not yield statistically significant findings at these institutionalized thresholds, many have feared the presence and consequences of unreported "failed" models and negative knowledge for the robustness of published sociological findings.

In medicine, especially clinical trials searching for a "treatment effect," these concerns have been formalized into the notion of statistical power. A study's statistical power represents its ability to demonstrate a causal relationship between two variables, given that such an association exists. In other words, power is a measure of that test's ability to avoid Type II errors or false negatives where a true signal is missed. A study with 80% power means that it has an 80% chance of resulting in a *p*-value of less than 5% if there was, in fact, an important underlying difference between the conditions studied. Study results will naturally be suspect if the study's statistical power is low. In a recent study, Katherine Button and colleagues found that the typical statistical power in neuroscience is only 0.21 or 21% (Button et al., 2013). Statistical power may be more difficult to calculate in sociological data analysis, where survey data is often partitioned differently to perform distinct statistical "experiments."

Concerns over inappropriate statistical inference have led to different approaches in other fields. Consider the field of high-energy physics (HEP) devoted to understanding the basic building blocks of matter. This field has remained highly organized for decades as it involves the arrangement of large teams around a few globally accessible particle accelerators. In 1960, the standard for discovery of a particle was 2σ, or two standard deviations from the center of a normal distribution and equivalent to a *p*-value of .05. As increased computational resources were allocated to analyzing a fixed number of accelerator experiments in the 1960's, this lead to a proliferation of published "discovery" of particles. The HEP community held a conference in

1968 at the University of Pennsylvania on meson spectroscopy that recognized this as a shift in statistical sensitivity and changed the community standard from 2σ to 3σ (Baltay & Rosenfeld, 1968), or from a *p*-value threshold of .05 to .003.[7] Because statistical significance is a result not only of study power but the likelihood of the hypothesis being evaluated and the ubiquitous bias favoring publication of novel discoveries, the HEP community instituted separate thresholds for "evidence of a particle" (3σ) and "discovery of a particle" (5σ—or a *p*-value threshold of .0000003)—the threshold used to evaluate the recent discovery of the Higg's Boson (ATLAS Collaboration, 2012; The CMS Collaboration, 2012).[8] With increased computational power and the recognition that negative knowledge typically remains unpublished, this same concern has been raised about sociological research (Freese, 2007), but the field is sufficiently decentralized that it has been very difficult to stage an organized shift in standards of significance.

Another approach to charges of unreported data mining is to report all tests estimated or comparisons made, but account for them explicitly when testing. In this broad approach, the standard has been Bonferroni correction, named after Carlo Emilio Bonferroni's inequalities, or upper and lower bounds on the probability of finite unions of events (Bonferroni, 1936), traceable in statistical usage to Olive Jean Dunn (1959, 1961). This approach is widely considered the simplest and most conservative method of controlling for the probability of making false discoveries (type 1) in multiple tests, or the familywise error rate. The approach naïvely assumes that the analyst is testing *k* independent hypothesis and so tests each hypothesis at $\tilde{\alpha} = {}^{1}\!/_{n} \times \alpha$. Most models tested by a researcher within a sequence of analyses, however, are not independent of one another. Alternative models typically contain variables that are

[7] Personal communication with Tom Witten, 2010.
[8] In a normal distribution, data is symmetrically distributed on both sides of the mean, but it is twice as likely for particle data to be in either the high or low tail than just the high tail, so the value is 0.0000003, or 1 in 3.5 million, rather than .0000006 (Lamb, 2012).

correlated, making the tests highly dependent. To impose such a harsh restriction on the threshold for identifying significance has contributed to the frequency with which researchers fail to disclose the number of models they have actually estimated.

More sophisticated approaches to multiple testing exist which increase statistical power, including multi-step (step-down, step-up) procedures to control the familywise error rate by accepting or rejecting hypotheses based on their ranked *p* values (Holm, 1979; Dunnett & Tamhane, 1992), or related approaches that explicitly control for the expected proportion of falsely rejected hypotheses or false discovery rate (Benjamini & Hochberg, 1995; Barber & Candes, 2014). These procedures have been embraced in the genomic sciences where the massive automated production of high-throughput experimental data has driven massive, simultaneous testing (Dudoit & Van der Laan, 2008). They have found much less appeal in the social sciences, even on massive Internet-based data sources. These approaches have found very little use with survey data, which are designed for a certain purpose and expensive to collect. As a result, the appearance of multiple testing has been systematically avoided rather than accounted for.

Consequently, the reader of a quantitative sociology article faces a knowledge asymmetry: only the author knows how many alternatives, if any, have been tested but found statistically insignificant and remain unreported (Gerber and Malhotra 2008; A. S. Gerber and Malhotra 2008; (Young, 2009). Should a reader facing this state of affairs discount confidence in everything he or she reads? And by how much? In other words, should the reader assume an even distribution of negative knowledge—knowledge held privately and unpublished by the authors—about what model specifications failed to reach statistical significance? These

judgments should ideally correspond to the articles' actual robustness, which is generally unknown.

There have been efforts to estimate the robustness of a small set of specific findings in sociology and related social and behavioral sciences. These are especially common where a published effect contradicts the findings or tenor of previous research. For example, consider the lively exchange regarding Tony Tam's 1997 article, "Sex Segregation and Occupational Gender Inequality in the United States," published in the *American Journal of Sociology* (Tam, 1997). Tam used the Dictionary of Occupational Titles (DOT) data (National Academy of Sciences, 1988) and purported to find that lower wages associated with high female composition occupations were not the result of systematic cultural devaluation of women's work, but rather underinvestment in specialized training. Paula England, Joan Hermsen and David Cotter responded, arguing that "using the same data used by Tam, we show that the addition of just one crucial control variable measuring occupations' demand for general education completely changes the results and restores the conclusion that there is a wage penalty for working in occupations with a higher % female" (England, Hermsen, & Cotter, 2000; see also Tam, 2000). Moreover, they argued that "virtually every study using DOT variables has included general educational development (GED)…. In contrast, [specific vocational preparation or] SVP, which Tam uses, measures vocationally specific preparation, whether obtained in school or on the job." While GED and SVP have a high correlation with each other (ρ>.7), they correlate differently with sex composition.[9] Such robustness exchanges are rare in the sociological literature.

[9] Tam responded that he was aware of the effect of GED, although he disputed its influence and disclosed that a reviewer had advised him to remove it from the paper (Tam, 2000).

One uncharacteristically large evaluation in the social sciences, the Reproducibility Project (https://osf.io/ezcuj/wiki/home/) represents a consortium of more than one hundred and fifty investigators, all attempting to reproduce findings from the 2008 issues of three journals including the *Journal of Personality and Social Psychology*. Laudable efforts like these require enormous effort by experts and are thus limited in scale. Here, we present a novel and relatively scalable approach that uses published models and original data, paired with minimally perturbed models and out-of-sample data. Using a sample of articles that use the GSS, we re-estimate these models after substituting a single, random variable with a close cognate; we also estimate original models on all available future waves of GSS.

Data and Methods

The General Social Survey (GSS) is a longitudinal, nationally representative survey of Americans conducted by the National Opinion Research Center (NORC) and designed to monitor and explain changes in attitudes, beliefs, and attributes across a wide variety of social spheres. From its first wave in 1972, the GSS has served as one of the premier data resources for social scientists, who have used it in more than 20,000 publications and perhaps even more classrooms (Gibson, 2013). NORC staff has periodically searched the academic literature for publications that use the GSS and for each publication recorded the years of the survey used, the variables used, and other metadata.[10] These efforts produced a database with metadata describing thousands of social science articles, books, and other publications. Andrew Abbott used these data in his article, "Seven types of ambiguity" in order to explore the diverse readings of an indicator capturing strength of religious attachment (Abbott, 1997).

[10] Until the mid-1990s these curation efforts included nearly 100% of the relevant publications. Budget limitations since that time have prevented NORC from curating but a portion of the discovered publications (Tom Smith, personal communication, 2014).

We supplemented these metadata with information describing in more detail the data analysis performed in each publication. For the subset of articles containing information on the variables used, a group of eight undergraduate coders located the original documents and supplemented the existing metadata by filling out a "survey" to collect additional pieces of information. This information included: (1) the method of data analysis used (i.e., single variable, bivariate or multivariate analysis; linear or nonlinear—e.g., logit analysis; predictive or nonpredictive—e.g., loglinear analysis); (2) whether the variables were treated as dependent, independent, or controls in the analysis; and (3) which variables—dependent or independent—were central to the article's research question. All articles were coded by at least three independent student coders. To evaluate the accuracy of the coders, we recruited a sample of authors associated with 97 of these articles to fill out the same survey and found moderate agreement (Cohen's κ) and correlation (Pearson's ρ) between the author and the majority of coder evaluator assessments with all but the "central" variables, where authors coded "central" variables much more selectively than coders. Because coders were not equally accurate and sometimes disagreed, we used a generative, probabilistic model to estimate coder accuracy (e.g., *Does Variable X operate as a dependent variable?*) and estimate a posterior probability for each variable coding (Rzhetsky, Shatkay, & Wilbur, 2009). See appendix and Table A4.2 for details.

These metadata enable us to approximate, but not necessarily replicate, each article's data analysis on the original GSS data. To do so we make several simplifying assumptions. First, we model the relationships between the variables using multiple linear regressions. Not all articles in our sample use multiple regression as their chief methodology. Nevertheless, regressions continue to be the modal quantitative method in sociology and are modal in our sample of

articles.[11] Second, to estimate these models, we use all of the GSS data in a given year and not a sub-sample of it, for example those observations belonging to a particular social group, as some articles do. Third, we use the following functional form. Each dependent variable Y_i is regressed on an intercept, all independent variables X^* and all control variables X°.

$$Y_i = \beta_0 + \sum \beta_j X_i^* + \sum \beta_k X_i^\circ + \varepsilon$$

The same right-hand-side specification is used for each dependent variable used in the publication. All variables are standardized by their standard deviations. This model is estimated on each year of GSS data separately[12]. Details of how the models were estimated and some illustrative examples of published papers, for which this estimation is a reasonable approximation, may be found in the Appendix (see Figure A4.4).

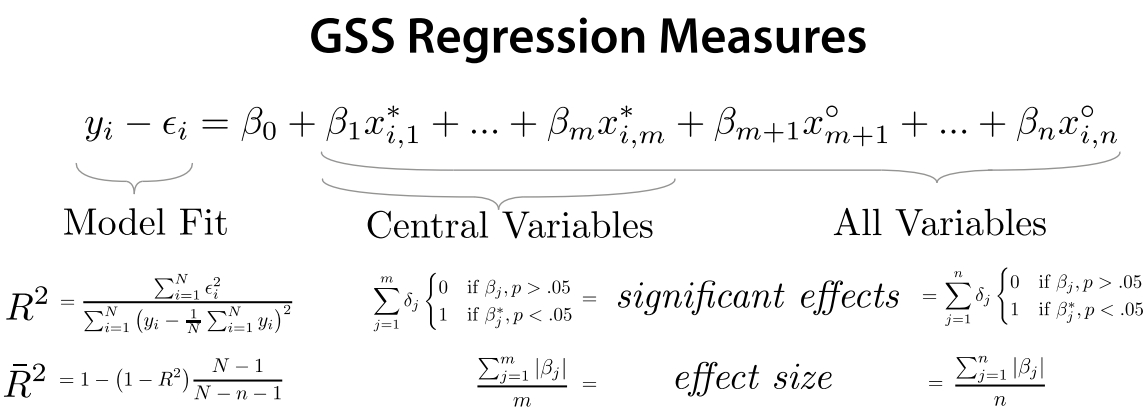
Measures of model fit

For each model, we record several outcomes that capture how well the model fits the data and the strength of the associations between dependent and independent variables (Figure 4.1). For goodness of fit, we record R^2 and adjusted R^2 or $\bar{R}^2$. For strength of associations between variables, we record the proportion of coefficients (ignoring the constant) that are statistically significant, average of the absolute values of the standardized coefficient sizes, and all of these outcomes for the subset of independent variables identified by coders as "central" to the analysis. Approximating the data analysis of articles in this way enables us to test the robustness of findings at time of publication and several decades afterward.

[11] Excluding from analysis articles that did not use linear regression did not materially affect our results.

[12] Each article thus may contribute several data points: a model for each dependent variable (model) and a point for each year of GSS on which each model was estimated, *i.e.* (num. of DVs) * (num. of GSS years).

Figure 4.1. Measurement strategy for capturing model fit and significance across original and perturbed model



Robustness at time of publication

We evaluate robustness at the time of publication by perturbing our approximation of published models in two ways: (1) by substituting one, randomly selected independent variable that was central to the analysis with a close cognate, and re-estimating the model; and (2) by estimating models on waves of the GSS that appeared immediately after publication.

To identify cognate variables suitable for substitution, we used the GSS subject index available on the NORC website[13]. The subject index divides all variables into hierarchical topical groups. For example, under the "Abortion" topic are five sub-topics, one of which, "Arguments pro and con" lists nine variables, such as "Importance of abortion issue to respondent." Given a variable, its "cognates" were defined as variables within the same (deepest) topical sub-group in the subject index that shared a correlation of at least 0.6 with the focal variable. Table 4.1a presents the most common *original variable → cognate variable* substitutions.

[13] http://www3.norc.org/GSS+Website/Browse+GSS+Variables/Subject+Index/. Accessed 9/5/2014.

Table 4.1a. Most common original-cognate variable substitutions.

Original variable	Cognate variable	Number of substitutions
EDUC	DEGREE	205
EDUC	SPEDUC	192
ATTEND	SPATTEND	153
EDUC	SPDEG	128
EDUC	MAEDUC	68
EDUC	PAEDUC	52
RACMAR	RACMAR10	31
DEGREE	SPDEG	28
WORDJ	WORDSUM	27
DEGREE	SPEDUC	26
PAEDUC	PADEG	25
DWELOWN	DWELLING	21
WORDSUM	WORDE	14
DEGREE	EDUC	14
PAEDUC	MAEDUC	9
PAEDUC	MADEG	8
PRES72	PRES68	8
LIBHOMO	LIBATH	8
COLATH	COLSOC	6
MAEDUC	PAEDUC	4

Definitions of these variables are provided in Table A4.1b in the Appendix.

An additional way in which we evaluate the robustness at the time of publication is by comparing models estimated on the last wave of GSS data used in each publication to those estimated on the GSS wave immediately following publication. This comparison is designed to limit the importance of changes in the social world, which are presumed to operate on time scales longer than one or two years. The newer wave of the data may be seen as additional draw of a sample from the original population.

Robustness to social change over time

In order to estimate how robust models are to a changing social world, we estimated models on the most recent year of data used in each article and compared their fit to those from estimates on each subsequent[14] year of the GSS. This time-series of changes allows us to observe how much the model fit degrades, if at all, with each additional year of social change. When possible (when a publication's variables exist in future GSS waves) we generate these trajectories of model fit for several decades. Additionally, we consider the very beginning of such a trajectory—the difference between a model's fit of the last year of the article's original GSS data and the fit on the *very next available year*[15]—as a perturbation of the relationship's robustness at time of publication. We assume that the social world only rarely changes substantially in so short a time. This perturbation in data complements the perturbation by substitution of a cognate variables described earlier as a signal of robust relationships near time of publication.

Results

This section presents measurements of the robustness of approximate published models at time of publication and in the subsequent decades. For each perturbation, the same set of outcomes was recorded: overall model fit (R^2 and adjusted-R^2 or $\bar{R}^2$), importance of central independent variables (proportion of these with p-values < 0.05 and average standardized coefficient sizes), and importance of all independent variables (proportion of these with p-values <0.05, and average standardized coefficient sizes).

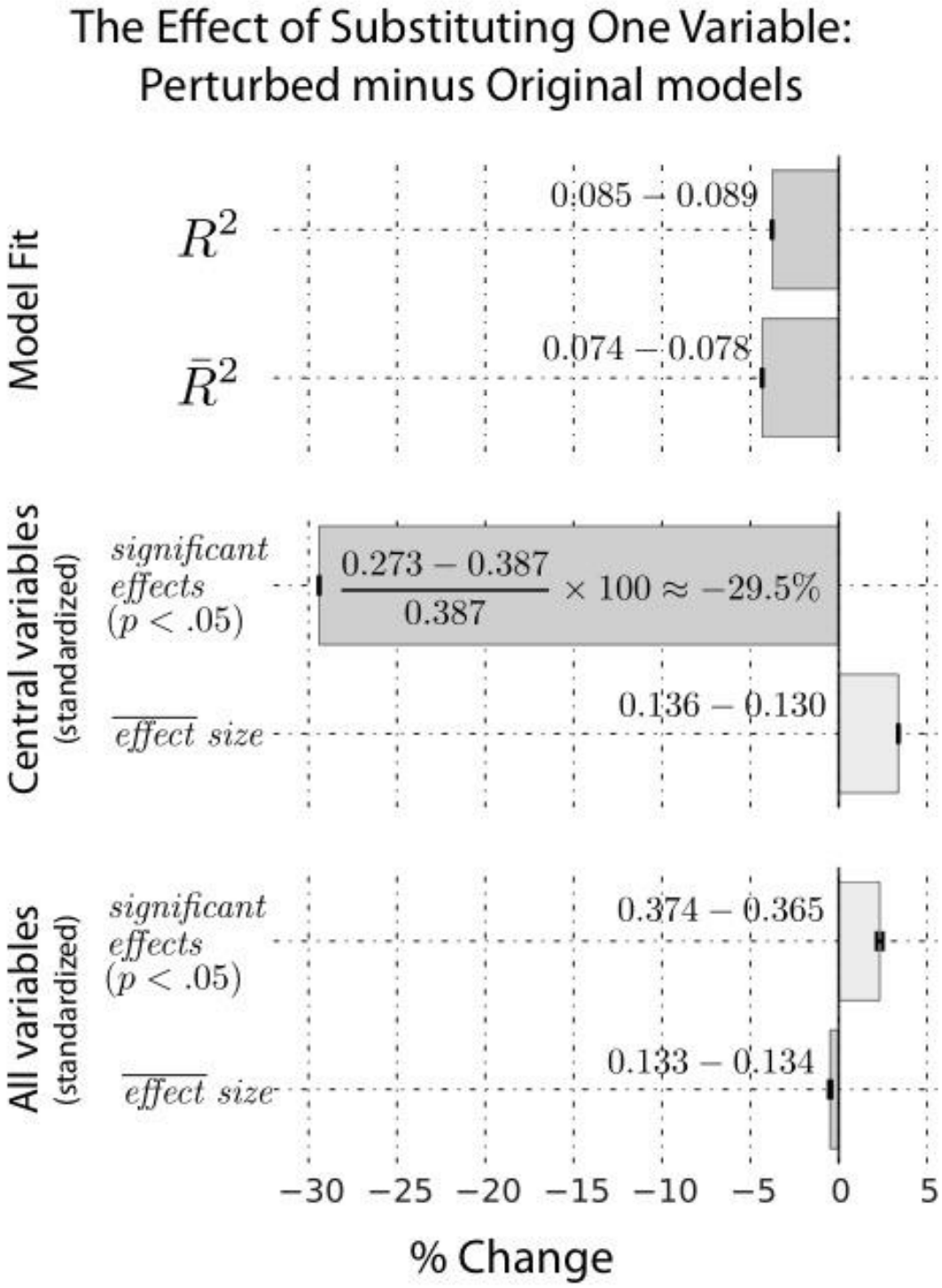
[14] Many variables in the GSS appear in just a few waves of the survey; only the hundred or so core variables reappear every year. Thus, models from many articles could not be estimated on future GSS years because subsequent data did not contain the necessary variables.

[15] The next suitable year depends on the variables used in the article and those available in subsequent GSS waves. For analysis of models estimated on last-year-used vs. next-available-year we did not consider articles where the next available year occurred more than 3 years later.

Robustness at time of publication: Cognate variable substitution

In order to evaluate how robust models are at time of publication, we first estimated our approximation of the original models, as described above (see Appendix). Then we perturbed these models by randomly selecting an independent variable central to the analysis with a close cognate. Robust models should be little affected by such a perturbation and should continue to fit the data equally well. On the other hand, models that are not robust, for example those in which a particular variable was chosen over its cognates because the *p*-value of only this coefficient fell under 0.05, should fit the data more poorly. Figure 4.2 below summarizes the outcomes of this experiment. The x-axis shows percent change in each outcome, while the raw outcomes, original minus perturbed, are printed to the right of each bar. The sample consists of models from 250 articles; the error bars extend to ± 2 robust clustered standard errors.

Figure 4.2. Change in model fit after the original model was perturbed by the substitution of one randomly selected central variable. One of the bars shows the calculation of percent change of mean outcome from the original to the perturbed models. The numbers next to the other bars show the mean outcomes of the perturbed and original models.



The top panel of Figure 4.2 shows percent change in R^2 and adjusted R^2 ($\bar{R}^2$). After perturbation, both of these measures decrease by 4-5%. The middle panel summaries associations between the dependent variable and those independent variables deemed central to the analysis (central IVs). The largest difference is in the proportion of these central IVs with statistically significant coefficients. After perturbation this quantity decreases by approximately 29%.

Figure 4.3. Shift in the distribution of significant central variables with the substitution of one, randomly selected variable.

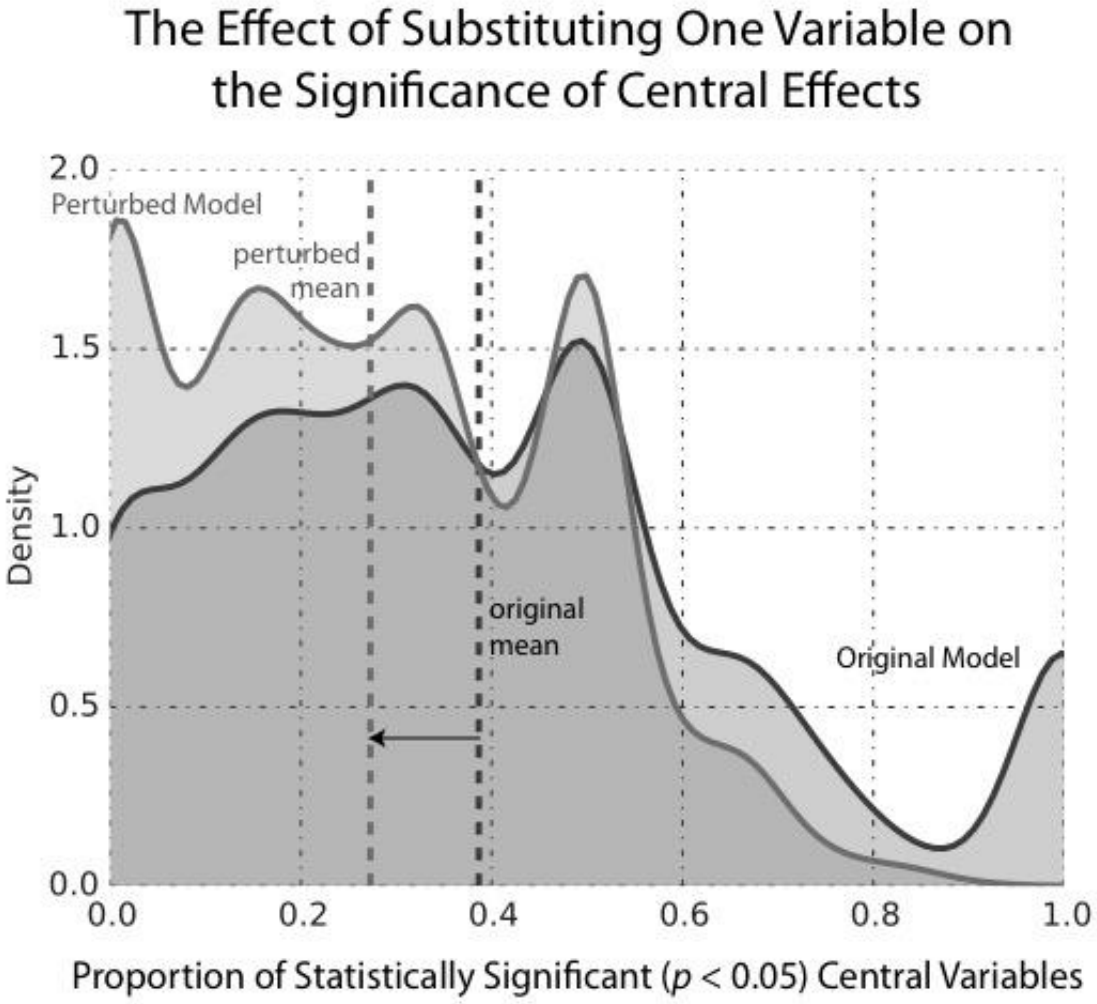


Figure 4.3 illustrates the density of the full distribution of significant central variables before and after perturbation by replacing a single, randomly selected central dependent variable with a close variable (see Table 4.1a).

The other outcomes change in the unexpected direction. For example, the standardized coefficient sizes of central variables increase by about 3%.

Robustness at time of publication: Data substitution

The second test of robustness at time of publication was performed by perturbing models through the substitution of data. By estimating them on the last year of GSS data used in the publication and on the very next available GSS year—the first "future" year[16]. The sample for this analysis draws on 398 qualifying articles. Figure 4.4 summarizes the consequences of this perturbation.

[16] Often the variables used in the article were not available in the very next GSS year but were available in a later wave. We included in our analysis articles whose variables were available in GSS data at most 3 years after the last GSS year the article used.

Figure 4.4. Change in model fit after the original model was re-estimated on data the next available year after publication. One of the bars shows the calculation of percent change of mean outcome from the original to the perturbed models. The numbers next to the other bars show the mean outcomes of the perturbed and original models.

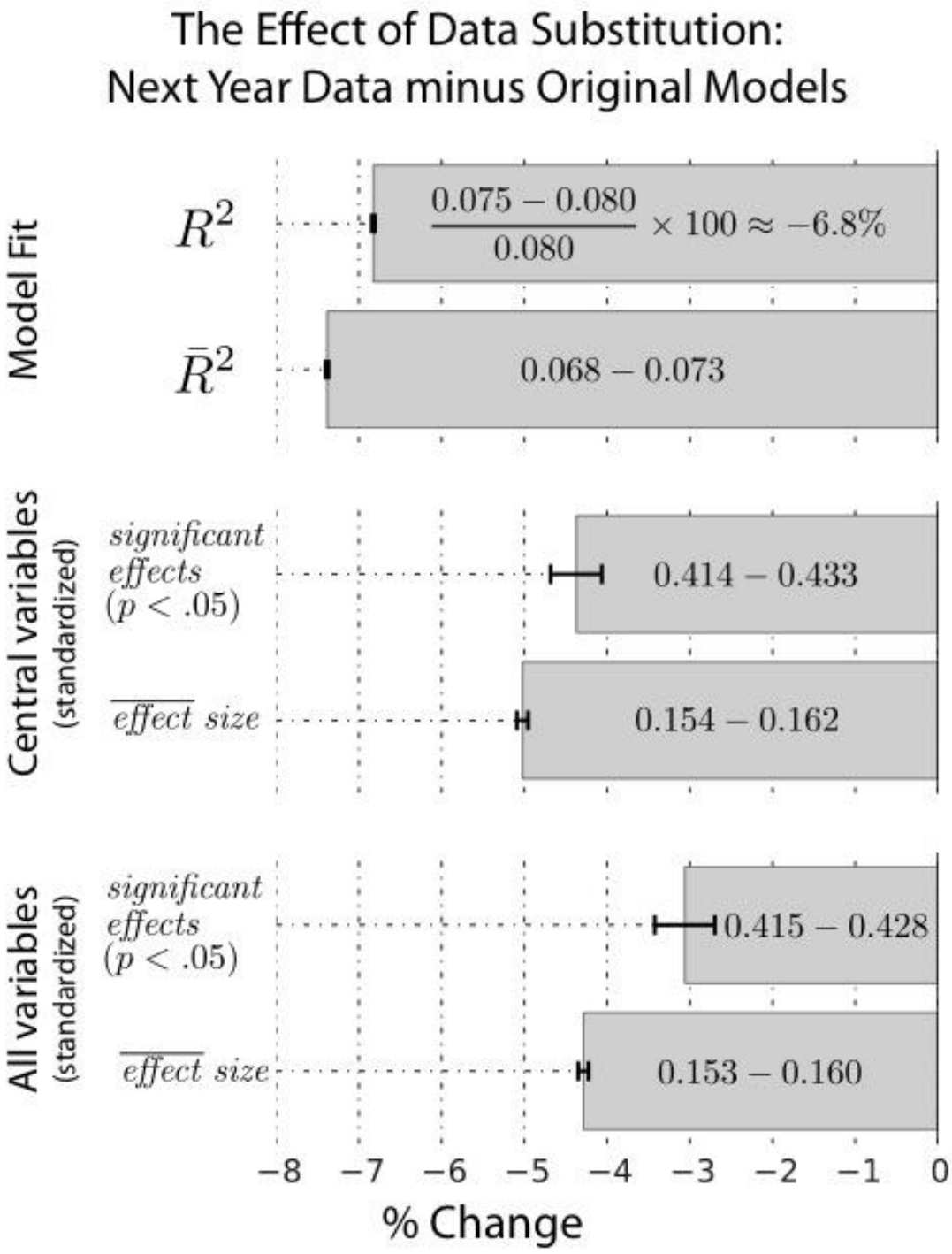
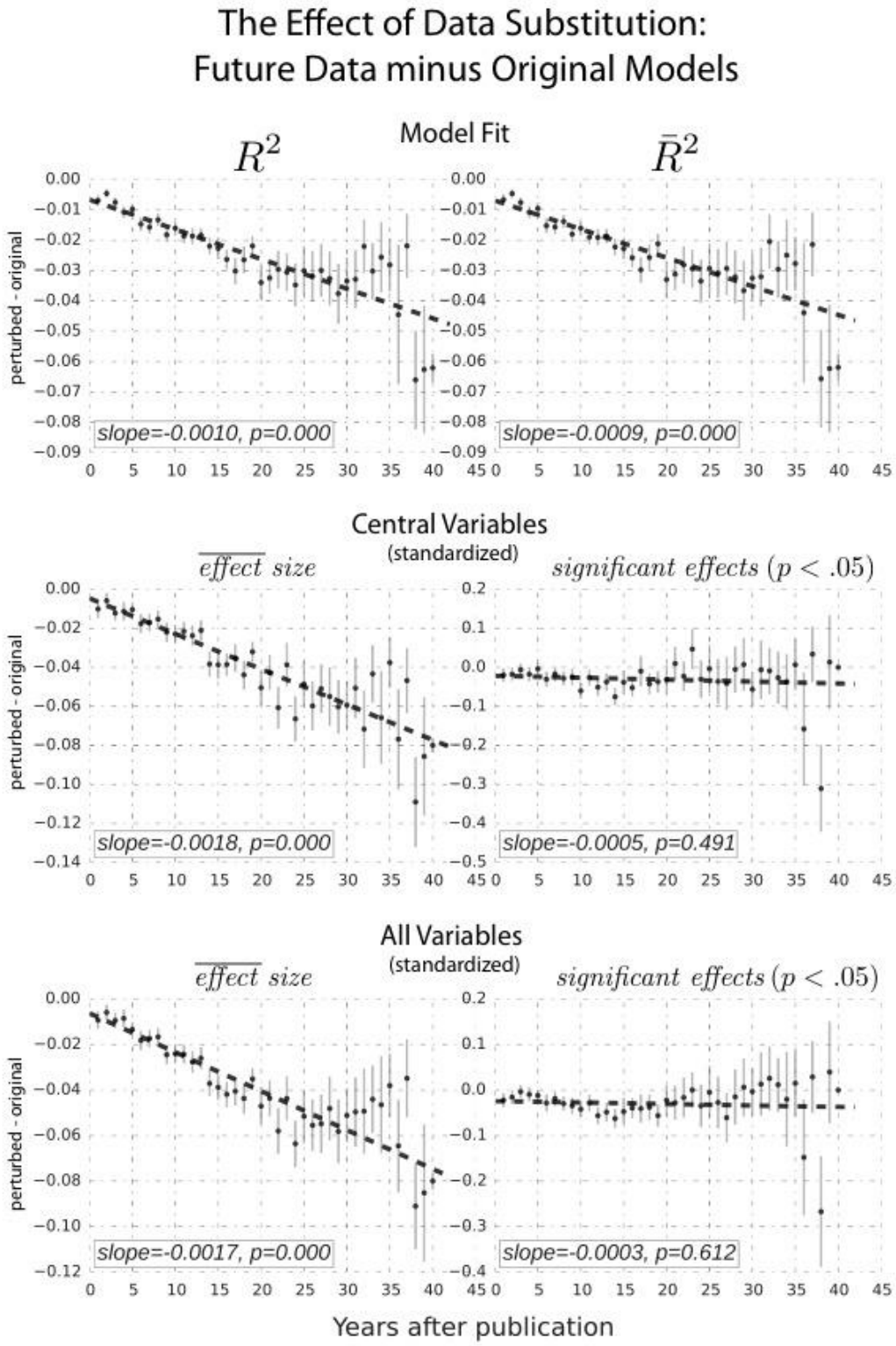


Figure 4.4 shows that R^2 and adjusted R^2 ($\bar{R}^2$) decrease by 6-8%. Standardized coefficient sizes of central and all independent variables decrease by 4 to 4.5%. Decreases in proportions of statistically significant coefficients dropped by 3 to 4%.

Robustness to Social Change

Lastly, we investigated the robustness of published findings to social change over time by estimating models on the last year of data used in the published articles and then comparing their fit to those same models estimated on each subsequent GSS year. Figure 4.5 presents these results. The *x*-coordinate of each dot is the number of years elapsed between the last GSS year used in a publication and some future GSS year used to re-estimate original models. The *y*-coordinate is the average difference between model outcomes when estimated on a new and the "original" year. The sample consists of 469 articles and bars represent ± 2 conventional standard errors. The dotted regression lines measure linear change in these differences over time, and their *p*-values are due to robust, clustered standard errors.

Figure 4.5. Change in model fit after the original model was re-estimated on data in each available year following publication.



Articles published relatively recently have less representation in the figure: they will not be found in the right side of each panel, as they have not existed long enough for decades to have passed since publication. This explains why the standard errors grow from left to right as the data become sparser. The top left panel displays the discrepancy between the original and perturbed R^2 over time. As one moves to the right on the *x*-axis, the time-gap between the last year of GSS used in the article and a future GSS year increases. Downward sloping lines in all panels suggests that all models become worse as the world changes, and old models fail to describe new realities.

Each additional year after the last year of GSS used is associated with 0.001 decrease in R^2. Over 10 years, this accumulates to a 0.01 difference; over 30 years the difference is 0.03. The typical R^2 in the present data sample is approximately 0.1, so after 30 years of social change, the R^2 decreases by 30% on average. R^2 and most other model outcomes show significant decline over time, indicating that models describe the data more and more poorly. The proportion of statistically significant coefficients for central and all independent variables drift downward, but do not significantly change over time; this apparent consistency may be the result of a small effect that is counterbalanced by gain in statistical significance due to the increase in the GSS sample size over time (see Appendix, Figure A4.3.).

It is instructive to consider the magnitude of these decreases in model fit to those caused by perturbing models with variable substitution and (relatively small) data substitution. Across all outcomes, the original model specifications describe GSS data more poorly 30 years after publication than perturbations to the model specification (or a small data perturbation). For example, as mentioned previously, R^2s decrease by about 30% after 30 years, but decrease only by about 7% when model specification is perturbed by a variable substitution.

Conclusion

This paper estimated the robustness of findings published in articles that use the General Social Survey (GSS). Previous research has questioned the robustness of findings like these because authors may have reported only those analyses that yielded "desired"—surprising, significant or newsworthy—results, and not the "negative knowledge" about which analyses failed such outcomes. Published findings may also fail to be robust due to social change. As the world changes, due to forces exogenous to published social science or in reaction to it, findings that once described the world accurately may cease to be so. To compare the effects of these two mechanisms on the quality of the GSS literature we extracted from the publications the statistical models and then "perturbed" them in two ways: (1) we tweaked the model specification by randomly substituting one important variable with a close cognate then re-estimated the model on each year of original data, and (2) we estimated the original model on newer waves of the GSS.

Previous efforts to estimate the robustness of the social scientific literature have been based on indirect methods and have generally yielded pessimistic results (J. P. Ioannidis, 2005; J. B. D. Long & Lang, 1992; Simmons et al., 2011; Simonsohn, Nelson, & Simmons, 2014; Young, 2009). In contrast, our method tested robustness relatively directly. The effects of the perturbations are generally small, indicating that published findings are relatively stable at the time of publication. Thus data-mining practices that produce unreported negative knowledge and weaken findings do not appear to be widespread in our sample. On the other hand, robustness decreases substantially over time and this may very well be the bigger concern for the reader of a sociological literature. The rate of "robustness lost" due to data-mining compares to that lost during about a decade of social change.

The method we demonstrate is particularly exciting because of its generality. In contrast to laudable but capital-intensive efforts to replicate published findings one-by-one, the present method scales. It requires few inputs: the models used in a corpus of articles and publically available data that, ideally, reoccurs periodically. These inputs already exist for a variety of popular longitudinal datasets. Websites where such datasets are hosted often provide several waves of the data and bibliographies of published research. We encourage and expect our method to be applied to other literatures. It is especially interesting to compare the results presented here to literatures in which researchers collect their own data, potentially on idiosyncratic population samples, and data sharing and re-use is not the norm.

The analyses presented here suffer from a number of limitations. First, and perhaps most crucially, the data include only articles published before 2006, and the data become sparser as we approach the present (see Figure A4.1 in the Appendix). As described earlier, concerns over robustness and reproducibility have been fueled in part by how easily statistical analyses may be used (and abused) with modern software. It is possible that such software has become even easier to use since 2005 and the increased facility may have engendered increased *p*-hacking.

Second, the corpus is composed of articles whose chief data resource was the GSS. The GSS is an immensely popular data resource in the social sciences; it is second only to the U.S. Census in the number of articles in which it has been used (Gibson, 2013). Nevertheless, research that uses the GSS forms only a small part of the social science literature, and it may be the case that GSS articles fail to represent the larger literature. For example, GSS articles may use the dataset's longitudinal nature specifically to study social behavior that changes over time. Moreover, consider the distinction between opinions, which can change relatively quickly, and fundamental social processes, which change more slowly. The GSS has been used to study both,

but it is unclear if either of these types of research dominates the literature. For example, authors may be interested in the change over time in Americans' attitudes toward abortion. In such a case it is no surprise that robustness of findings over time will be relatively low, because the inquiry was undertaken *because* the outcome changes.

It is also important to investigate the social characteristics of the producers and outlets of weak findings. For example, do lower-tier journals publish weaker findings than the top-tier journals?

Lastly, many assumptions were needed to successfully estimate thousands of models from hundreds of articles and these may be found in the Appendix. Despite these many limitations, we believe that our high-throughput method of evaluation can be usefully extended, and that the findings we present provide a rich trace of the partially obscured process through which sociological findings emerge, achieve publication and, with sufficient time, dissolve into social history.

Discussion

In recent years, science studies scholars have trained their sights on a subject close at hand: the social sciences. Sociological research (Leahey, 2008b), evaluation (Michèle Lamont, 2009b, 2012), and reporting (Franco et al., 2014) practices have rightly become not only things we perform but subjects we attempt to understand and reflect upon. Much remains to be learned about what it is we do as a knowledge culture (Knorr-Cetina, 1999). In this paper we estimate the robustness of published sociological findings—widely visible ice caps of knowledge—and from the results attempted to deduce the potentially larger underwater iceberg of negative knowledge that researchers produce through performing unsuccessful statistical analyses along the way. Many fear data-mining or *p*-hacking is rampant and undermines the robustness of much

sociological literature. We find that the robustness of the literature, in the period that our data covers, does not suffer greatly from this practice. Sociologists do not appear to engage in extensive data-mining. In fact, it is interesting to consider what would happen if sociologists systematically (and perhaps a-theoretically) mined social data. Would the knowledge generated from such a research practice differ from the current? Would unexpected relationships be discovered?

The second issue we raised in this article was the impact of social change on the literature. The social world changes and some published findings no longer hold. We present a systematic analysis of how much social change weakens the published literature and find this source to affect the literature more than unreported data-mining. Every ten years, the socio-cultural world changes sufficiently to overwhelm the effects of selective publication and unpublished negative knowledge.

With the emergence of passively collectible "big data" from social media, our exclusive reliance on survey sources for this information may attenuate. Social media can also reveal rich traces of human attitudes, opinions and behavior, and insofar as data from these sources can be inexpensive to harvest, our anxiety with "discovering" hypotheses in data may diminish proportional to the ease with which we can identify new, out-of-sample data on which to test them. In conjunction with our findings that social change over the march of time erodes more published findings than *p*-hacking, we may recommend that as a field we engage in more *reported* data mining, rather than less.

Appendix 4

Table A4.1b: Variable definitions

Variable	Definition
EDUC	Respondent's formal education: 0 (none) – 20 (8 years past HS)
DEGREE	Respondent's highest degree: 0 (less than HS) – 4 (graduate)
ATTEND	How often do you attend religious services? 0 (never) – 8 (many times a week)
SPATTEND	How often spouse attends religious services: 0 (none) – 8 (more than a week)
SPDEG	Spouse's highest degree: 0 (less than HS) – 4 (graduate)
MAEDUC	Mother's formal education: 0 (none) – 20 (8 years past HS)
PAEDUC	Father's formal education: 0 (none) – 20 (8 years past HS)
RACMAR	Favor law against racial intermarriage? (yes/no)
RACMAR10	Favor law against racial intermarriage 10 years ago? (yes/no)
WORDJ	Vocabulary test: identify words similar to word J: (correct/incorrect)
WORDSUM	Number of words correct in a vocabulary test?
PADEG	Father's highest degree: 0 (less than HS) – 4 (graduate)
DWELOWN	Does respondent own or rent home?: 1 (own), 2 (rent), 3 (other)
DWELLING	Dwelling type: 1-10 (types, e.g. trailer, apartment house)
WORDE	Vocabulary test: identify words similar to word E: (correct/incorrect)
PRES72	If voted, did you vote for McGovern or Nixon?
PRES68	If voted, did you vote for Humphrey, Nixon, or Wallace?
LIBHOMO	Allow homosexual book in library? (remove/not remove)
LIBATH	Allow anti-religious book in library? (remove/not remove)
COLATH	Allow anti-religionists to teach? (yes/no)
COLSOC	Allow socialist to teach? (yes/no)

Coding of the data

Student researchers affiliated with the National Opinion Research Center, which administers the GSS, searched the academic literature for articles using the GSS and coded these articles for the presence of specific GSS variables. These data are publically available on the GSS website. We subsequently and independently employed 6 undergraduate student researchers with sociological and methodological training to locate these articles and (a) confirm whether the variables purportedly used in each article were indeed employed and (b) identify the role of each variable in the statistical analysis: dependent variable, independent variable, central variable, and/or control variable. These classifications (dependent, independent, central, control) were not

treated exclusively. For example, a variable could be coded as both dependent and independent in a paper that predicts an outcome and then uses this outcome to predict another variable. An example of a set of models well characterized by our re-estimation approach is shown in Figure A4.3 (Mary F. Fox & Firebaugh, 1992), where the variable "CONSCI" or confidence in science is predicted by (regressed on) gender in one model, alongside a number of comparison models where gender predicts confidence in other institutions. Other pieces of metadata regarding GSS variable analyses were also recorded by the students, as illustrated in Table A4.1.

Table A4.1. Metadata associated with variables linked to each article from the sample

Metadata	Description
Dependent variables	Variables to be explained
Independent variables	Variables treated as explanatory
Central variables	Variables treated as of primary interest
Control variables	Variables treated as a control
Mode of data analysis	Model of the relationship between the variables (e.g., linear regression)
GSS years used	Years of the GSS used in the publication
GSS years future	Years of the GSS that follow the last year used and contain all relevant variables

All articles were then reread and coded by a balanced set of three student coders using a 6 choose 3 design, such that all possible 3-coder-subsets (20) coded an equal number of articles.[17] Coding was performed through a website that allowed students access to the digital article. Table A4.2 lists the average pairwise Cohen's κ and Pearson's ρ between all coders for each class of variable assignments. The Cohen's κ values for agreement on dependent and independent variables are in the .4 to .5 range, described as "moderate" by Landis and Koch (1977) or "fair to good" by Fleiss (1981, p. 218). Agreement for central and control variables was slightly lower, between .3 and .4, what Landis and Koch call "fair" (1977), despite criticism that standard thresholds of acceptability are not appropriate (Gwet, 2012), for example, because κ grows with

[17] One coder dropped out before completion of the task and so we introduced a new coder in their place. Our estimates of accuracy, described below, included all seven coders.

the number of codes. In our analysis, only two codes are possible (0/1) and so the scores are naturally somewhat lower than they would be otherwise.

To evaluate the validity of the student codes, we recruited a sample of authors associated with 97 of our published articles to fill out the same online survey and we uncovered moderately high agreement between author and student coder assessments, for all except the "central variables category." Dependent and independent variable Cohen's κ scores were .37 and .54, respectively, with control variables lower at .28. The central variable coding relationship was virtually unrelated (-.06), explained by the fact that authors only determined 35% of the variables in a paper as "central", while students determined 75%. This is likely because authors viewed what was "central" by their *ex ante* expectations and research design, whereas coders only had access to the *ex post* narrative of the analysis, and which variables and findings had *become* central in the final document. This deviation makes our analyses more conservative than they otherwise would have been, because "central" variable substitutions in our model perturbation experiments, described in the following sections, in many cases simply consists of what the author would have considered a "control" variable substitution, and so possibly involved less intensive search on the part of the author/analyst. We believe that these student assessments of variable "centrality" are also semantically meaningfully, insofar as these variables had become part of the published narrative available to audiences.

Table A4.2. Agreement between core coders and authors

Item of model coded	Dependent Variable	Independent Variable	Central Variable	Control Variable
Average pairwise Cohen's κ for 6 student coders	.48	.45	.27	.33
Average pairwise Pearson's ρ for 6 student coders	.52	.47	.32	.39
Average value for authors (1=yes, 0=no)	.47	.24	.35	.82
Average value for student coders (1=yes, 0=no)	.36	.31	.75	.94
Cohen's κ for author & mode of student coders	.37	.54	-.06	.28
Pearson's ρ for author & mode of student coders	.39	.55	-.09	.28
Cohen's κ for author & discrete posterior	.41	.45	-.02	.30
Pearson's ρ for author & continuous posterior	.51	.54	-5.7	.33

Because not all coders coded items with equal accuracy, and because "don't know" was an optional answer, leading to potential ties, we used a generative, probabilistic model to estimate the maximum *a posteriori* probability (MAP) prediction that an item's code is true, which integrates over the estimated accuracy of coders, assuming only that the entire population of coders is slightly more often right than wrong. The model ("Model B") is based on a simple underlying generation process that directly accounts for the probability that coded values are correct (Rzhetsky et al., 2009). For each coded value j, a set of parameters, denoted γ_j, represents the probability that each coded value is correct. For the i^{th} coder ($i = 1, 2, \ldots, 6$), we introduce a matrix of probabilities, denoted $\lambda^{(i)}_{x|y}$, that defines the probability that she assigns code x (e.g., Dependent variable) to a GSS variable with correct annotation y. For a perfect coder, the matrix $\lambda^{(i)}_{x|y}$ would equal the identity matrix and her vote would count most toward the total. For a coder that always codes incorrectly—a "troll"—her matrix $\lambda^{(i)}_{x|y}$ will have all its value off the diagonal and will only minimally influence the posterior.

Table A4.2 shows that on average, Cohen's κ and Pearson's ρ between author codes and posterior estimates are comparable to those gathered from mode (popular vote) of the student coders. In some cases ("dependent variables" and "control variables") they are higher. Because these posterior break ties between the coders, they allow us to use all of the data available for analysis. We took the highest MAP estimate (the discrete posterior) for each code. Those variables coded "dependent" were used as *dependent* variables in our analysis; those coded "central" were considered *central*; and those coded "independent," "central," or "control" were used as *independent* variables.

Data Description

The following figures describe the publications included in the data sample, and the size of the GSS over time.

Figure A4.1 below displays the number of articles per year in the data sample. It is important to re-emphasize that the data do not include articles published after 2005, despite the existence of hundreds of such articles. Budget limitations have prevented the National Opinion Research Center, which administers the GSS, to fully urate the continuing flow of relevant publications after the mid-1990s (Tom Smith, personal communication, 2014).

Figure A4.1. Number of articles per year in the sample (n=1525).

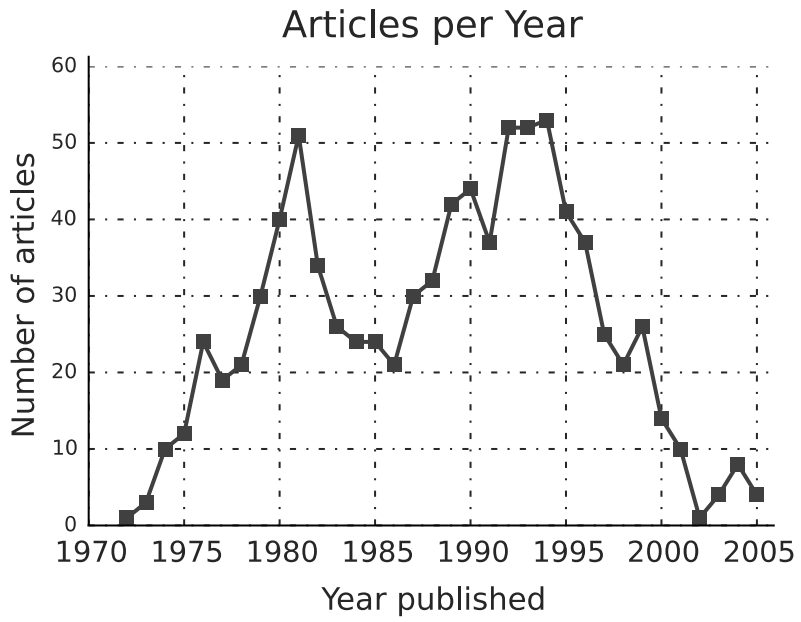


Figure A4.2 below displays time-trends in the average number of dependent and independent variables per article published in a given year.

Figure A4.2. Number of dependent and independent variables per article over time. (3-year moving averages; 1525 articles)

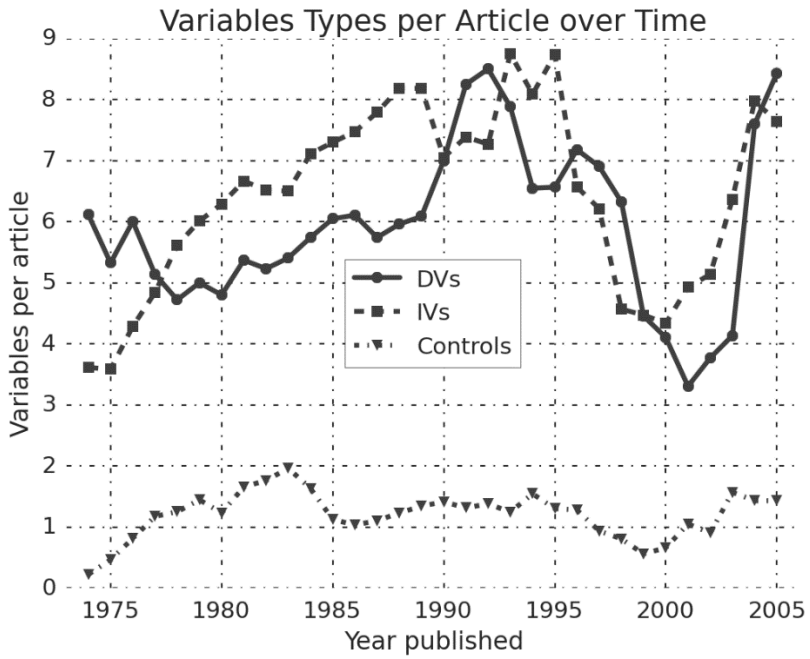
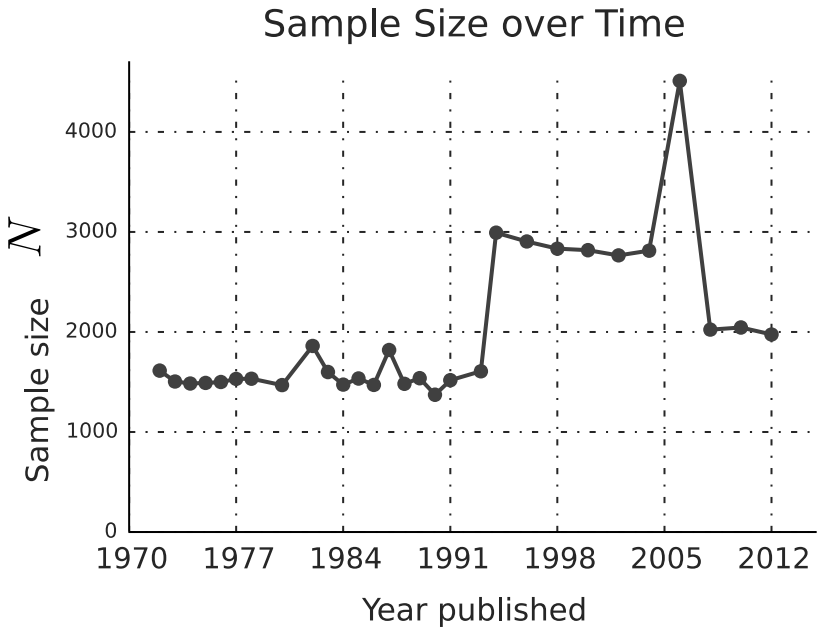


Figure A4.3 illustrates how the sample size of the GSS has changed over time. The numbers of persons sampled has increased (non-linearly) over time. This growth in the GSS makes some of our estimates – especially of proportions of statistically significant effects (coefficients) over time – conservative because p-values from t-tests on coefficients shrink with increasing sample size.

Figure A4.3. GSS sample size over time. Note that our data do not include articles published after 2005.

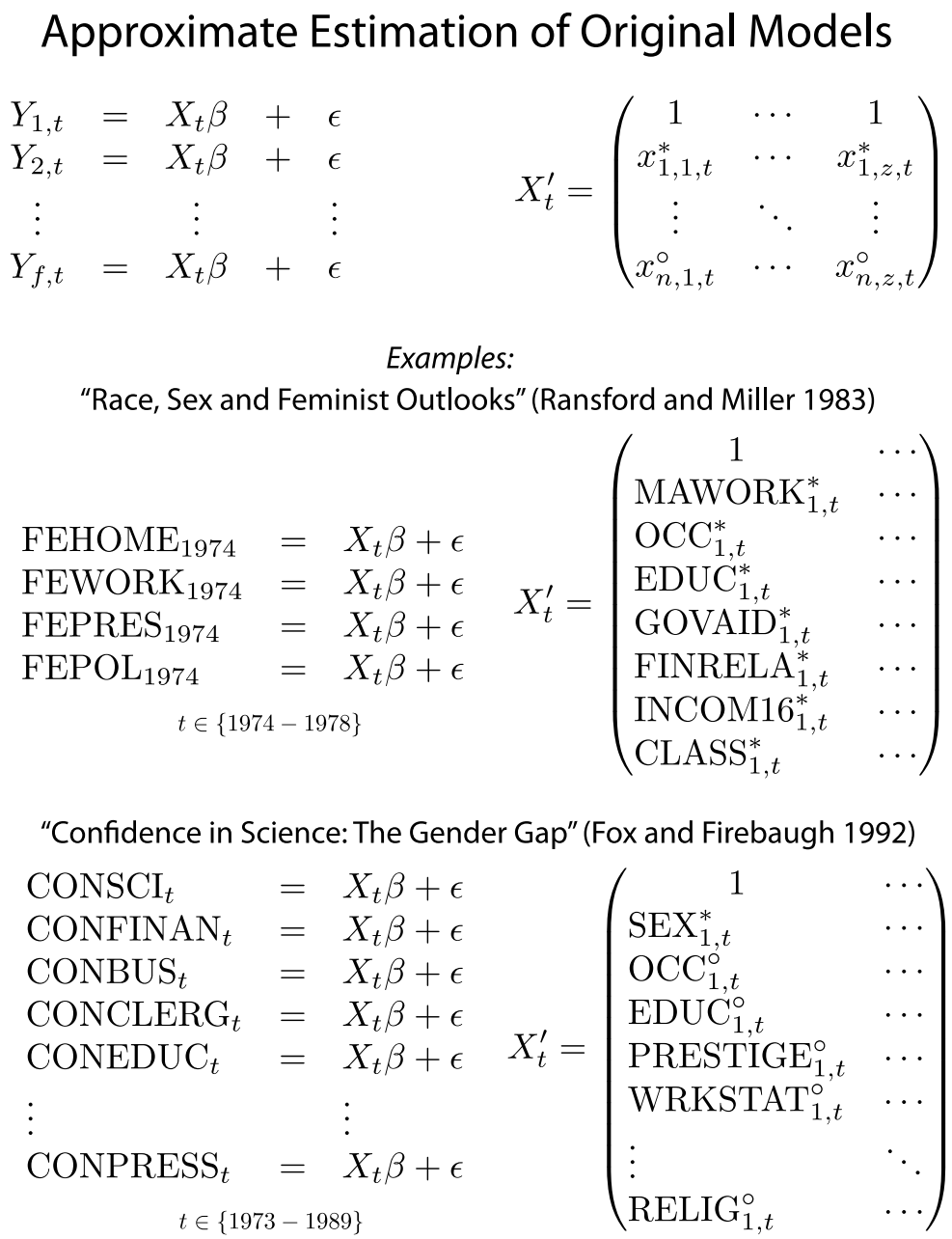


How models were estimated

The relevant data for each publication consists of the GSS years used, dependent variables, independent variables, control variables, and those independent variables central to the publication's analysis and argument, the so-called "central variables." Forty-five articles did not have data on years of GSS used; GSS years used was imputed for these articles as all GSS years prior to the year of publication in which all of the article variables were present. Each (standardized) dependent variable was regressed simultaneously on all of a publication's

(standardized) independent variables and controls, and the regression was estimated separately on each year of GSS data. Figure A4.4 below illustrates this approach and provides examples of two publications from the data sample.

Figure A4.4. Schema used to approximate replication of original models with examples from (Ransford & Miller, 1983) and (Mary F. Fox & Firebaugh, 1992).



Several assumptions were required to successfully estimate these model specifications.

Variables types. First, many variables in the GSS are categorical and necessitated special treatment. We examined the 300 most commonly used variables and, if they were categorical, identified how many levels were possible. Categorical *independent* variables with more than 15 levels, e.g. DENOM (specific protestant denomination) and OCC (census occupation code), were not included in the regressions. Regressions in which the *dependent* variable was categorical with more than 2 levels were also skipped. Variables coded on Likert scales were treated as continuous.

Missing values. Many sociological articles do not impute missing values and simply drop records with any missing information. We chose instead to impute values because our approach of regressing dependent variables on *all* independent variables within the article, even if there are 30 such variables, made it likely that there would be few, if any, fully complete cases. We imputed missing values for each variable using that variable's mean or, in the case of categorical variables, the mode. This naïve choice for imputation was made to best approximate real research practices in the sociological literature, where the most prevalent strategy was no imputation at all, and if imputation was performed, it is usually with the mean. Imputing using more sophisticated methods, including regression of a missing variable on all others, multiple imputation (Rubin, 2004), or the use of low-rank matrix models to simultaneously impute based on column and row similarity (Udell, Horn, Zadeh, & Boyd, 2014), significantly changes the means of the variables and, while perhaps closer to "sociological truth", such data lies further from the data authors of GSS publications actually searched. We should expect that performing our robustness analyses on these "improved" samples would show decreased difference between "original" and "perturbed" models because both were estimated on perturbed data. We find that

this to be the case, with most effects becoming nonsignificant, but retaining the same direction of those presented.

Clustered standard errors

The articles in our data usually use models with several dependent variables, several independent variables, and estimate these models on several years of data (see Figure A4.2). In our analyses we used several simplifying assumptions, including (a) models are estimated separately on each year of data and (b) models separately regress each dependent variable on all independent variables. Thus, in many cases a single article provided several data points (one for each dependent variable and one for each year of data used). In such cases, observations are clustered (by article) and may give rise to correlated errors, which tend to make t-tests and coefficient estimates from OLS regressions yield inappropriately small standard errors. To test the robustness of our results we performed significance tests using clustered (by article) standard errors.

CHAPTER 5. AMPLIFYING THE IMPACT OF OPEN ACCESS: WIKIPEDIA AND THE DIFFUSION OF SCIENCE*

Abstract

With the rise of Wikipedia as a first-stop source for scientific knowledge, it is important to compare its representation of that knowledge to that of the academic literature. Here we identify the 250 most heavily used journals in each of 26 research fields (4,721 journals, 19.4M articles in total) indexed by the *Scopus* database, and test whether topic, academic status, and accessibility make articles from these journals more or less likely to be referenced on Wikipedia. We find that a journal's academic status (impact factor) and accessibility (open access policy) both strongly increase the probability of its being referenced on Wikipedia. Controlling for field and impact factor, the odds that an open access journal is referenced on the English Wikipedia are 47% higher compared to paywall journals. One of the implications of this study is that a major consequence of open access policies is to significantly amplify the diffusion of science, through an intermediary like Wikipedia, to a broad audience.

Introduction

Wikipedia, one of the most visited websites in the world[1], has become a destination for information of all kinds, including information about science (Heilman & West, 2015; Laurent & Vickers, 2009; Okoli, Mehdi, Mesgari, Nielsen, & Lanamäki, 2014; Spoerri, 2007). Given that so many people rely on Wikipedia for scientific information, it is important to ask whether and to what extent Wikipedia's coverage of science is a balanced, high quality representation of the knowledge within the academic literature. One approach to asking this question involves looking

* Co-authored with Gracu Lu and Eamon Duede and forthcoming in the *Journal of the Association for Information Science and Technology.* Reprinted here with permission from Wiley.

[1] http://www.alexa.com/siteinfo/wikipedia.org. Accessed 2015-06-15.

at references used in Wikipedia articles. Wikipedia requires all claims to be substantiated by reliable references[2], but what, in practice, are "reliable references?"

An intuitive approach is to examine whether the sources Wikipedia editors use correspond to the sources scientists value most. In particular, within the scientific literature, a journal's status is often associated, albeit problematically (Seglen, 1997), with its impact factor. If status within the academic literature is taken as a "gold standard," Wikipedia's failure to cite high impact journals of certain fields would constitute a failure of coverage (Samoilenko & Yasseri, 2014), while a high correspondence between journals' impact factors and citations in Wikipedia would indicate that Wikipedia does indeed use reputable sources (P. Evans & Krauthammer, 2011; Nielsen, 2007; Shuai, Jiang, Liu, & Bollen, 2013).

Yet high impact journals often require expensive subscriptions (Björk & Solomon, 2012). The costs are, in fact, so prohibitive that even Harvard University has urged its faculty to "resign from publications that keep articles behind paywalls" because the library "can no longer afford the price hikes imposed by many large journal publishers" (Sample, 2012). Consequently, much of the discussion of open access focuses on the consequences of open access for the scientific community (Van Noorden, 2013). A lively debate has arisen on the impact of open access on the scientific literature, with some studies showing a citation advantage (Eysenbach, 2006a, 2006b; Gargouri et al., 2010; "The Open Access Citation Advantage Service," n.d.) while other find none (Davis, 2011; Davis, Lewenstein, Simon, Booth, & Connolly, 2008; Gaulé & Maystre, 2011; Moed, 2007).

[2] https://en.wikipedia.org/wiki/Wikipedia:Verifiability. Accessed 2015-06-15.

Apart from a rather unclear impact on the scientific literature, open access journals may have a tremendous impact on the diffusion of scientific knowledge *beyond* this literature. To date, this potential of open access policies has been a matter chiefly of speculation (Heilman & West, 2015; Trench, 2008). Previous research has found that open access articles are downloaded from publishers' websites more often and by more people than closed access articles (Davis, 2010, 2011), but it is currently unclear by whom, and to what extent open access affects the use of science by the *general public* (Davis & Walters, 2011). We hypothesize that Wikipedia, with more than 8.5 million page views *per hour*[3], diffuses scientific knowledge to unprecedented distances (Joe Wass, 2015)and that diffusion of science through it may relate to accessibility in two ways. By referencing findings from paywall journals, Wikipedia distills and diffuses these findings to the general public. On the other hand, Wikipedia editors may be unable to access expensive paywall journals[4], and consequently reference the easily accessible articles instead. For example, Luyt and Tan's (Luyt & Tan, 2010) study found accessibility to drive the selection of references in a sample of Wikipedia's history articles. In this case Wikipedia "amplifies" open access science by broadcasting its (already freely accessible) findings to millions. This "amplifier" effect may thus constitute one of the chief effects of open access.

Correspondence between academic and Wikipedia statuses

This article tests both the distillation and amplifier hypotheses by evaluating which references Wikipedia editors around the world use and do not use. In particular we study the correspondence between journals' status within the scientific community (impact factor) and their accessibility (open access policy) with their status within Wikipedia (percent of a journal's

[3] http://stats.wikimedia.org/EN/TablesPageViewsMonthly.htm. Accessed 2015-06-16.

[4] Wikipedia has recently partnered with major publishers to provide editors access to some paywall literature: http://en.wikipedia.org/wiki/Wikipedia:The_Wikipedia_Library. Accessed 2015-09-02.

articles referenced in Wikipedia). It is important to note that an observed correspondence may be evinced by a variety of mechanisms besides the aforementioned accessibility. First, the status ordering of academic journals as measured with impact factors may have only a tenuous relationship with the importance and notability – considerations of special relevance to Wikipedia[5] – of the published research. Citations, and therefore impact factors, are in part a function of the research field (Seglen, 1997), and may be affected by factors as circumstantial as whether a paper's title contains a colon (Jamali & Nikzad, 2011; Seglen, 1997). Second, the academic status ordering results from the objectives of millions of scientists and institutions, and may be irrelevant to the unique objectives of Wikipedia. Wikipedia's key objective is to serve as an encyclopedia, not a medium through which scientists communicate original research[6]. Relative to the decentralization of the scientific literature, Wikipedia is governed by explicit, if flexible, policies and a hierarchical power structure (Butler, Joyce, & Pike, 2008; Shaw & Hill, 2014). Apart from a remark that review papers serve Wikipedia's objectives better than primary research articles, Wikipedia's referencing policies generally pass no judgment over which items within the scientific literature constitute "the best" evidence in support of a claim[7]. Wikipedia's objectives and explicit, centrally accessible, policies differ from the decentralized decisions that produce status orderings within the scientific literature and do not imply that the two status orderings should correspond. Indeed, if editors are not scientists themselves they need not even be aware that journal impact factors exist[8]. On the other hand, despite the well-worn caveats, prestigious, high-impact journals may publish findings that are more important to both academics and Wikipedia's audience. In fact, a Wikipedia editor's *expectation* that the truly

[5] http://en.wikipedia.org/wiki/Wikipedia:Notability. Accessed 2015-06-11.
[6] http://en.wikipedia.org/wiki/Wikipedia:Five_pillars. Accessed 2015-05-29.
[7] http://en.wikipedia.org/wiki/Wikipedia:Identifying_reliable_sources#Scholarship. Accessed 2015-05-29.
[8] Citation metrics often influence the ranking of academic search results and may thus promote high impact journals without searchers' knowledge.

important research resides within high-impact journals may be enough to predispose them to reference such journals. Second, little is known about editors of science-related articles (West, Weber, & Castillo, 2012); they may be professional scientists with access to these high-impact journals, resulting in both the *motivation* and *opportunity* to reference them.

Previous research

Wikipedia references and academic status

The first large-scale study of Wikipedia's scientific references was performed by Finn Arup Nielsen (Nielsen, 2007). Nielsen found that the number of Wikipedia references to the top 160 journals, extracted from the *cite-journal* citation templates, correlated modestly with the journal's *Journal Citation Reports* impact factor. This implication that Wikipedia preferentially cites high impact journals is delicate in part because the data used in the study included only a subset of journals with references that appear in Wikipedia, not journals that were and *were not* referenced. It is possible, albeit unlikely, that an even larger number of prestigious journals, made invisible by the methodology, are never referenced on Wikipedia at all, weakening the correlation to an unknown degree, or that the referenced journals are simply those that publish the most articles (see Nielsen 2007: Fig. 1). Shuai, Jiang, Liu, and Bollen (2013) also found modest correlations when they investigated a possible correspondence between the academic rank of computer science papers, authors, and topics and their Wikipedia rank.

The altmetrics movement has also explored Wikipedia as non-academic venue on which academic literature makes an impact (ALM, Fenner, & Lin, 2014; "altmetrics," n.d.; Priem, 2015). Evans and Krauthammer (P. Evans & Krauthammer, 2011) examined the use of Wikipedia as an alternative measure of the scholarly impact of biomedical research. The authors correlated scholarly metrics of biomedical articles, journals, and topics with Wikipedia citations

and, in contrast to other studies, included in some of their analyses a random sample of journals never referenced on Wikipedia. The authors also recorded a journal's open access policy but, unfortunately, do not appear to have used this information in analyses.

Open access and the Web

The rather voluminous literature on open access has focused primarily on effects on the academic literature[9]. There is some debate on the size and direction of open access effects. Some evidence demonstrates that open access articles gain a citation advantage (Eysenbach, 2006a, 2006b; Gargouri et al., 2010; "The Open Access Citation Advantage Service," n.d.), while other evidence shows no such effect (Davis, 2011; Davis et al., 2008; Gaulé & Maystre, 2011; Moed, 2007). Regardless of the impacts on scientists in developed nations, increased accessibility through open access does yield benefits to scientists from developing nations (Davis & Walters, 2011; J. A. Evans & Reimer, 2009).

The promise of open access for disseminating scientific information to the world at large has gained much less attention (Davis & Walters, 2011; Trench, 2008); for an exception see (Heilman & West, 2015). Yet, more and more of the world turns to the Web for scientific information. For instance, as early as 1999 a full 20% of American adults sought medical and science information online (Miller, 2001). What's more, one who actively seeks such information within the academic literature will quickly discover that, despite the paywalls, many important and impactful research articles are made freely available by their authors or third parties (Björk, Laakso, Welling, & Paetau, 2014; Wren, 2005a). This is to say nothing of the fact that science may also be disseminated through distillation of its findings into venues like

[9] This literature has grown to thousands of items and is impossible to summarize fully. See (Craig, Plume, McVeigh, Pringle, & Amin, 2007; Davis & Walters, 2011) for two reviews of parts of this literature.

Wikipedia or science-centric websites and blogs so that, here too, the impact of open access may be limited. While full texts of the most impactful literature are, at least nominally, behind a paywall (Björk & Solomon, 2012), do Wikipedia's editors consult these texts? If they cite them in Wikipedia, have they consulted the full texts beyond a freely available abstract before referencing? If the academic literature is any guide, referenced material is sometimes consulted rather carelessly (Broadus, 1983; Rekdal, 2014). In short, the current understanding of the relationship between open access and the general public in the literature is limited at best (Davis & Walters, 2011).

Shortcomings and present contribution

In addition to the role of accessibility, a number of substantive and methodological shortcomings remain. First, it is unclear if professional scientists edit Wikipedia's science articles. As we will show below, a preponderance of paywall references would suggest, albeit indirectly, this to be the case[10]. The scant existing evidence indicates that science articles are edited by people with general expertise, relative to the more narrow experts of popular culture articles (West et al., 2012). Second, most previous studies have completely ignored the articles that are never referenced on Wikipedia, thus sampling on the dependent variable. The only notable exception, (P. Evans & Krauthammer, 2011), treated the unreferenced articles outside the main analytic framework. While the framework treated (referenced) articles or journals as the unit of analysis, the unreferenced articles and journals were treated as a homogeneous group.

This study extends existing work in three chief ways. First, it models the role of accessibility (open access status) on referencing. Second, it covers *all* major research areas of

[10] As corroborating evidence consider the list of Wikipedia editors by [self-reported] degree lists more than 1000 users with PhDs: http://en.wikipedia.org/wiki/Category:Wikipedians_with_PhD_degrees. Accessed 2015-09-02.

science by observing rates at which Wikipedia references nearly 5,000 journals, accounting for nearly 20 million articles. Third, it treats unreferenced articles in the same analytic framework as those referenced. Yet the study is not without its own limitations, which are outlined more fully in the discussion section. Chief among these are that article-level characteristics are operationalized by the characteristics of the publishing journal. For example, the accessibility of articles is operationalized by their journal's open access policy, when, in fact, free access to many paywall articles exists through sanctioned or unsanctioned file-sharing (Björk et al., 2014; Wren, 2005b). Thus, any observed advantage of open access referencing may be biased downward, i.e. an underestimate of the true effect (see the Conclusion for a discussion of measurement error).

Data and Methods

Data sample

Journal data

Our analysis uses journal-level data from thousands of journals indexed by *Scopus*. Indexing over 21,000 peer-reviewed journals and with more than 2,800 classified as open access, *Scopus* is the world's largest database of scientific literature[11]. We obtained information on the 250 highest-impact journals within each of the following 26 major subject areas[12]: *Agricultural Sciences, Arts and Humanities, Biochemistry and General Microbiology, Business Management and Accounting, Chemical Engineering, Chemistry, Computer Science, Decision Sciences, Earth and Planetary Sciences, Economics and Finance, Energy Sciences, Engineering, Environmental Sciences, Immunology and Microbiology, Materials Sciences, Mathematics, Medicine, Neurosciences, Nursing, Pharmacology, Physics, Psychology, Social Science, Veterinary*

[11] http://www.elsevier.com/online-tools/scopus/content-overview. Accessed 2015-01-24.
[12] The subject area "general" was excluded because it contained only four journals, all of which were cross-listed with other top-level categories.

Science, Dental, Health Professions. Assignment of journals to these broad subject areas is not exclusive; many journals fall into more than one category. As a result of cross listing, the list of candidate journals was less than 6500. The final data consisted of 4721 unique journals, 335 (7.1%) of which are categorized by the Directory of Open Access Journals as "open access."

Journals were also categorized more narrowly using the more than 300 "All Science Journal Classification" (ASJC) subject codes[13], e.g. Animal Science and Zoology, Biophysics, etc. These narrow codes were used to identify journals that address similar topics and thus indicate whether the journal is at risk for reference *vis-à-vis* demonstrated demand. Journals with at least one narrow subject code in common were considered "neighbors" and if at least one of these neighboring journals has been referenced the original "ego" journal was considered to be at risk for reference as well. Journals with no demonstrable demand were excluded from analysis. As an example, consider the journal *Science.* It is listed under (ASJC) subject code 1000 – general science. Other journals with this code – the "neighbors" of *Science* – are *Nature, PNAS,* and *Language Awareness*. *Language Awareness* is cross-listed under 5 others subject codes.

Impact factor was measured by the 2013 SCImago Journal Rank (SJR) impact factor. SJR correlates highly with the more conventional impact factor but takes into account self-citations and the diverse prestige of citing journals (González-Pereira, Guerrero-Bote, & Moya-Anegón, 2010; Leydesdorff, 2009). Table 5.1 displays the 15 highest SCImago impact journals, calculated with citations data available up to 2013.

[13] http://info.sciencedirect.com/scopus/scopus-in-detail/content-coverage-guide/journalclassification. Accessed 2015-06-03.

Table 5.1. 15 highest-impact journals within *Scopus* according to SCImago impact factor (2013).

Journal	Impact factor (SCImago2013)
CA - A Cancer Journal for Clinicians	45.894
Reviews of Modern Physics	34.830
Annual Review of Immunology	32.612
Cell	28.272
Annual Review of Biochemistry	27.902
Quarterly Journal of Economics	25.168
Nature Genetics	24.052
Nature Reviews Genetics	23.813
Nature Reviews Molecular Cell Biology	23.593
Chemical Reviews	23.543
Nature	21.323
Acta Crystallographica Section D: Biological Crystallography	20.717
Advances in Physics	20.349
Annual Review of Cell and Developmental Biology	19.686
Annual Review of Neuroscience	19.662

English Wikipedia data

References in the English Wikipedia were extracted from the 2014-11-15 database dump of all articles. We parsed every page and following (Nielsen, 2007) extracted all references that use Wikipedia's *cite journal* template. Since it allows editors to easily include inline references that are automatically rendered into an end-of-article bibliography, this template is the recommended way for editors to reference scientific sources in Wikipedia[14]. In all, there were 311,947 "cite-journal" tags in the English Wikipedia. An exploratory analysis of the 49 largest non-English Wikipedias can be found in the Appendix.

[14] Editors may also reference articles in other ways, for example by providing in-line links. We focus on the "cite-journal" template for three reasons. First, it shows clear intent to reference. Second, it has been used in previous research. Lastly, Ford and colleagues (Ford, Sen, Musicant, & Miller, 2013) found that "<ref>" tags were used most often to reference sources, and the "cite-journal" templates on which we focus are nested within such <ref> tags.

Matching *Scopus* journals to Wikipedia references

We checked each of the referenced journal names on Wikipedia against a list of *Scopus*-indexed journal names and common *ISI* journal name abbreviations. Of the 311,947 *cite-journal* tags, 203,536 could be linked to journals indexed by *Scopus*. Many of these references were non-unique, whereas our outcome of interest is whether articles from a journal are referenced on Wikipedia at all, not how many times. Therefore, to ensure that the counts for each journal included only unique articles, we distinguished articles by their DOIs and, if an article's DOI was not available, we used the article's title. *Scopus' coverage* of the output of various journals varies widely; our counts included only those articles published within the years of *Scopus* coverage.

In the end we matched 32,361 unique articles (and 55,262 total references) to our subset of *Scopus* journals (top 250 in each research field). 2,005 of the top *Scopus* journals are never referenced on the English Wikipedia. In most cases observed "journal names" that did not match to journals in *Scopus* were not academic journals but popular newspapers and magazines. Table A5.1 in the Appendix displays the 20 most frequently referenced sources that we were unable to link to *Scopus*. The top 3 non-science sources are *Billboard*, *National Park Service*, and *Royal Gazette*. However, efforts to match Wikipedia references to *Scopus* were imperfect, and the list also includes a handful of academic journals, including *The Lancet*.

Journal vs. article level unit of analysis

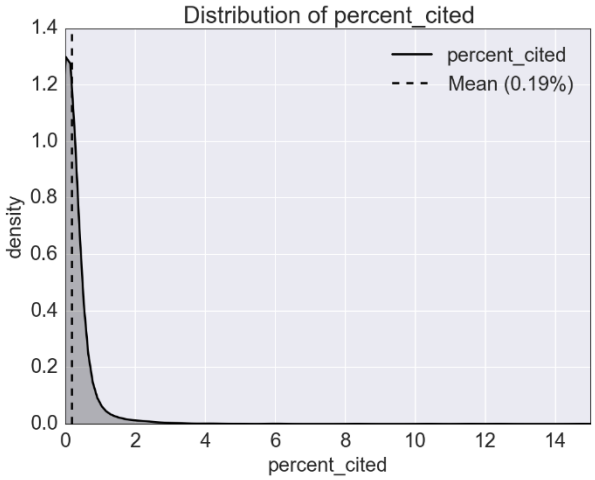
We chose to take journals instead of individual articles as our unit of analysis for several reasons. First and most important, accessibility of articles, the focal point of this study, was measured at the journal level by whether the journal is or is not open access. Switching the unit of analysis to individual articles would have simply assigned the same value of accessibility to

all articles from a particular journal. Second, while article-level citations are an attractive, finely grained metric, a journal's impact factor is also designed to capture citation impact, albeit more coarsely. The general topic of any given article is also well captured by the host journal's *Scopus*-assigned topic(s). Lastly, the matching of Wikipedia journal title strings to *Scopus* required some manual matching and these efforts were more practical at the level of thousands of journals instead of hundreds of thousands of articles.

percent_cited **and other variables**

We present some of our results in terms of *percent_cited* -- the *percent of a journal's articles that are referenced on Wikipedia*. An equivalent interpretation of this journal-level metric is the *probability that a given article from a journal is referenced on Wikipedia.* Figure 5.1 illustrates the distribution (kde) of *percent_cited*.

Figure 5.1. Distribution (kde) of *percent_cited* of 4774 journals.



As Figure 5.1 demonstrates, the vast majority of journals that scientists use are referenced on the English Wikipedia very little: on average 0.19% of a journal's articles are referenced[15].

[15] 2,005 (out of 4,721) journals are never referenced at all (*percent_cited* = 0).

As mentioned above, the academic status of journals was measured using (SCImago) impact factors. To limit the influence of the few journals with uncommonly high impact factors the impact factor variable was (natural) log-transformed when used in the models. Figure 5.2 displays the distribution of impact factor and log-impact factor; to aid visualization only journals with impact factor <=15 are shown.

Figure 5.2. Distribution (kernel density estimate) of impact factor and ln(impact factor). To aid visualization, impact factor > 15 is not displayed.

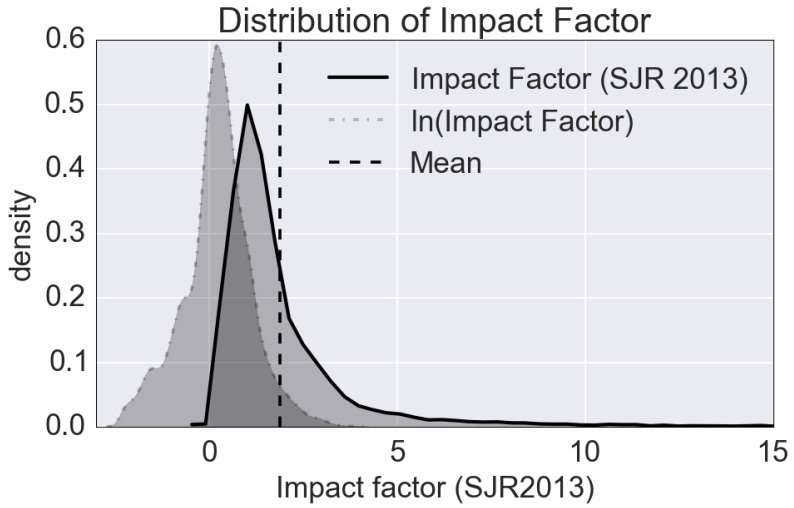


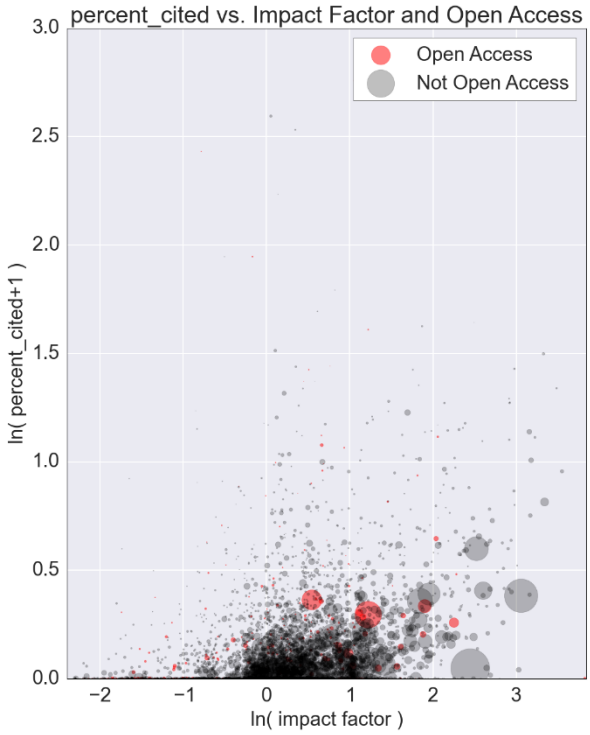
Table 5.2 presents the summary statistics for key variables: *percent_cited*, impact factor, ln(impact factor), and open access. Additionally, analyses use dummies for the 26 subject categories, e.g. psychology 0 or 1).

Table 5.2. Descriptive statistics of key variables.

Variable name	**Mean**	**Std.**	**Min**	**Max**
percent_cited	0.193%	0.545	0%	14.7%
impact factor	1.89	2.47	0.100	45.9
ln(*impact factor*)	0.212	0.909	-2.30	3.83
open_access	7.1% O.A.	----	0	1

Lastly, Figure 5.3 displays a scatter plot of the key dependent variable, *percent_cited*, versus impact factor and open access.

Figure 5.3. Scatter plot of journals' *percent_cited* vs. impact factor and open access policy. Dots are scaled by the total amount of articles published by each journal (and indexed by *Scopus*). Open access journals are shown with red dots.



The scatter plot appears to show a modest relationship between a journal's impact factor and *percent_cited*, the percent of its articles referenced on Wikipedia, especially when considering journal size (dot size). The next section analyzes these relationships statistically.

Results

We first present results of English Wikipedia's coverage. We ask, does Wikipedia draw equally on all branches of science? Next we focus on the role played by a journal's status and accessibility in predicting Wikipedia references. An exploratory analysis of references in the 49 largest *non-English* Wikipedias can be found in the Appendix.

Coverage

Figure 5.4 below summarizes which branches of the scientific literature the English Wikipedia draws upon. The left panel shows the number of articles published by the top 250 journals in each field. The right panel shows the percent of those articles that are referenced at least once in the English Wikipedia.

Figure 5.4. English-language Wikipedia's coverage of academic research. Each field's candidate literature (left) consists of the articles published by the 250-highest impact journals within the field and indexed by *Scopus*. The right panel shows the percent of these articles referenced on Wikipedia.

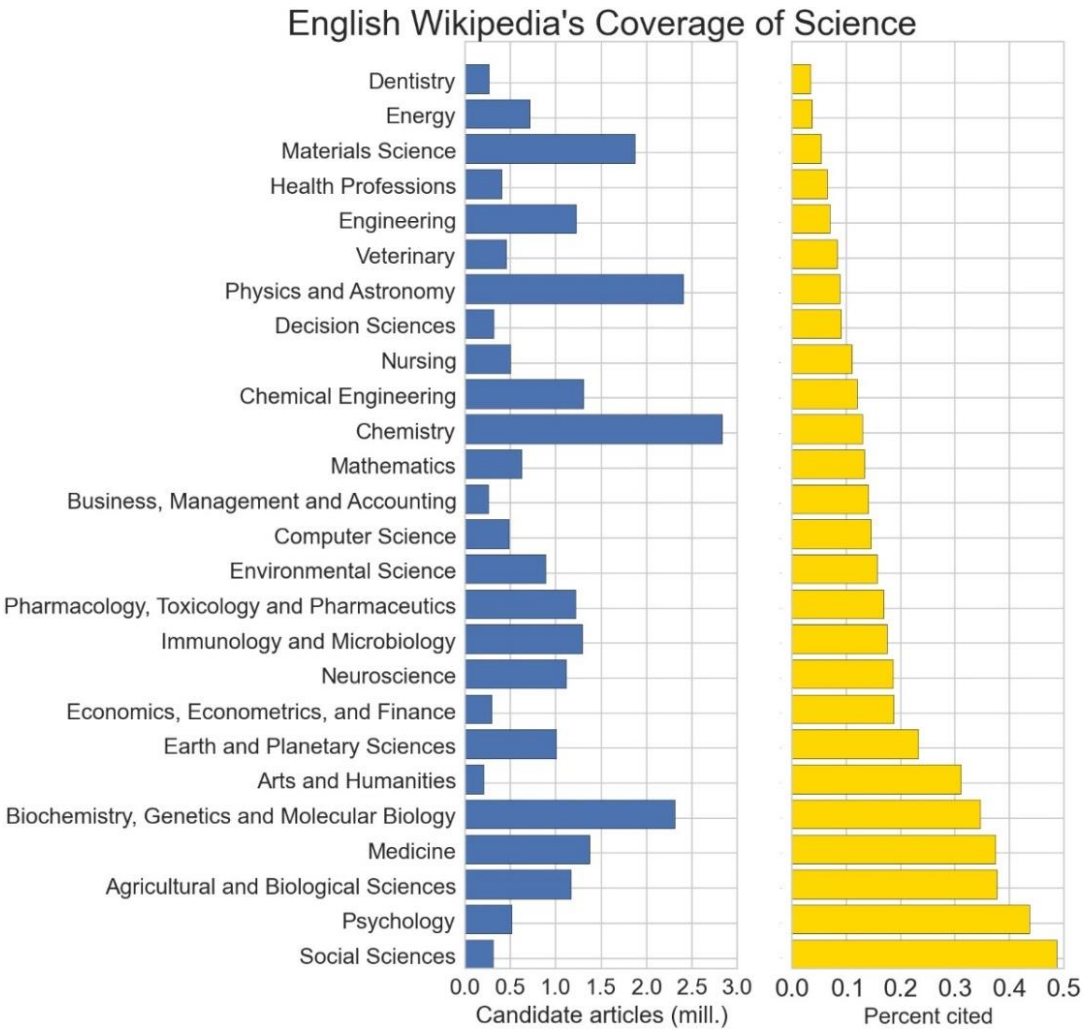


Figure 5.4 indicates that the coverage of science, as measured by the use of references, is very uneven and limited across scientific fields (Samoilenko & Yasseri, 2014). The social sciences represent a relatively small candidate literature but a relatively large portion of this literature is referenced on the English Wikipedia (0.4 – 0.5%). At the other end of the spectrum, dentistry, also a relatively small literature, is rarely referenced (< 0.05%). The ordering of disciplines by *percent_cited* does not engender a simple explanation. For example, such an ordering does not appear correlated with traditional distinctions like hard vs. soft science, or basic vs. applied. This finding is echoed by Nielsen (2007), who found that "computer and Internet–related journals do not get as many [references] as one would expect if Wikipedia showed bias towards fields for the 'Internet–savvy'". The highly uneven referencing across disciplines suggests that discipline should be controlled for in any statistical model, as is done below.

Status and accessibility

We now present results from an intuitive statistical model that predicts the probability *p* that an article from a given journal will be referenced given that journal's characteristics. The data-generating process is assumed to be a binomial process: each journal *i* publishes n_i articles and each of these articles is at risk p_i of being referenced in Wikipedia, where p_i depends on the journal. The probability that a journal *i* has *k* of its n_i articles referenced in Wikipedia is thus

$$Pr(y_i = k \mid n_i, p_i) \sim \binom{n_i}{k} p_i^k (1 - p_i)^{n_i - k}$$

. p_i is assumed to be a (logit) function of the journal characteristics $\boldsymbol{x_i}$'s (e.g. impact factor): $ln(\frac{p}{1-p}) = \mathbf{x}\beta$, where β are the parameters to be estimated. The model just described is commonly used for proportional outcomes: it embeds the

familiar logistic regression within a binomial process. This model is known as a generalized linear model (GLM) of the binomial-logit family (Hardin & Hilbe, 2012).

Table 5.3 below displays estimates from this model of how journal characteristics are related to its *p*, probability of referencing, fitted to the English Wikipedia.

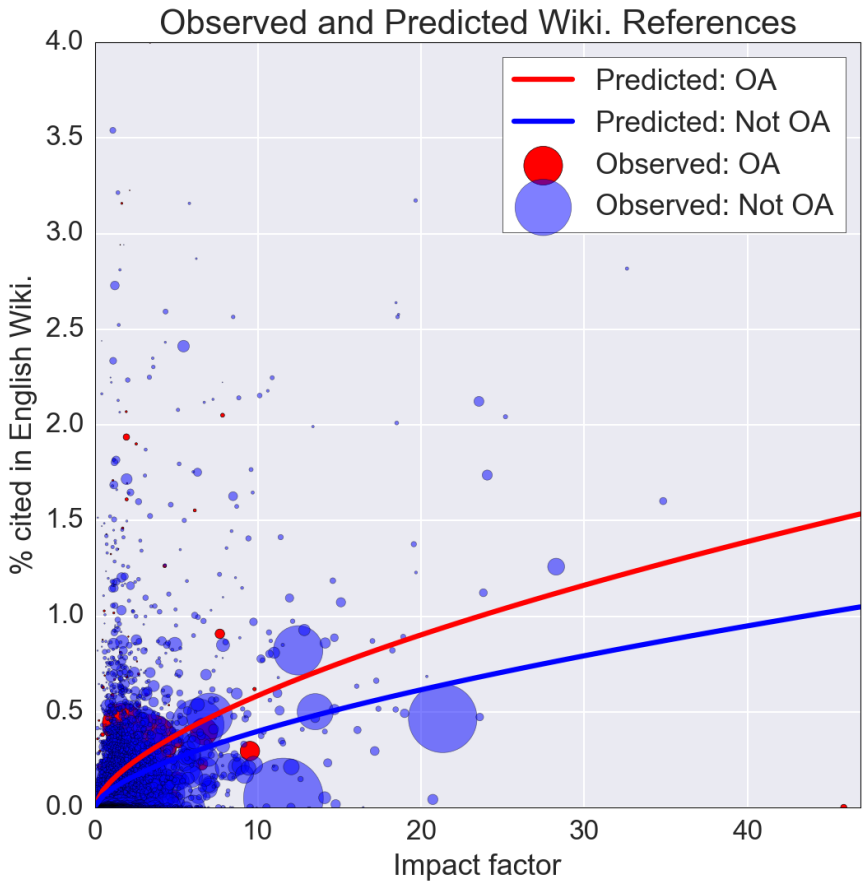
Table 5.3. Estimates from the GLM estimated on English Wikipedia reference data. Variables with statistically insignificant odds ratios are not bolded. *n*= 4720, d.f.=28.

Variable	**Odds ratio**	**95% C.I.**	**P-value**
open_access	1.471	1.406, 1.539	0.000
log_sjr2013	1.879	1.852, 1.906	0.000
ag_bio_sciences	2.292	2.210, 2.377	0.000
arts_hum	1.836	1.689, 1.996	0.000
biochem_gen_mbio	1.059	1.030, 1.090	0.000
bus_man_acct	0.714	0.638, 0.799	0.000
chem	1.004	0.962, 1.048	0.863
cheme	0.968	0.912, 1.027	0.282
cs	0.991	0.916, 1.074	0.831
decision_sci	0.957	0.844, 1.084	0.489
dental	0.520	0.422, 0.614	0.000
earth_plan_sci	1.515	1.446, 1.587	0.000
econ_fin	1.106	1.010, 1.210	0.030
energy	0.551	0.487, 0.642	0.000
engineering	0.507	0.471, 0.545	0.000
envi_sci	0.743	0.703, 0.787	0.000
healthpro	0.787	0.696, 0.891	0.000
immu_micro_bio	1.114	1.065, 1.165	0.000
materials	0.640	0.598, 0.685	0.000
math	0.716	0.664, 0.772	0.000
medicine	0.660	0.642, 0.679	0.000
neuro	1.033	0.986, 1.081	0.168
nursing	1.206	1.101, 1.313	0.000
pharm	1.481	1.409, 1.556	0.000
phys	0.629	0.599, 0.660	0.000
psyc	2.628	2.504, 2.760	0.000
socialsci	1.357	1.283, 1.436	0.000
vet	0.898	0.807, 1.000	0.048

The column of odds ratios indicates how the odds of referencing change with unit changes in the independent variables. For indicator variables, e.g. open access, these ratios are interpreted as the increase in odds when the indicator is true. For example, the odds that an article is referenced on Wikipedia increase by 47% if the article is published in an open access journal.

To interpret these results in terms of probabilities rather than odds ratios we must evaluate the model at particular values of the variables. Figure 5.5 displays the observed and predicted references for a range of values of impact factor and *open_access*. The indicator variables designating particular disciplines are set at their modes (0).

Figure 5.5. Observed (dots) and predicted (solid lines) English Wikipedia references. Red dots designate open access journals. The dot size is proportional to the number of articles the journal published.



The figure demonstrates that a journal's impact factor has positive and asymptotic effect on the percent of its contents referenced in the English Wikipedia (*percent_cited*). Open access journals (red dots) are relatively uncommon, but these journals are referenced more often than paywall journals of similar impact factor. For example, in our sample of psychology journals, open access journals have an average impact factor of 1.59, while closed access journals have an average impact factor of 1.77. Yet in the English Wikipedia, editors reference an average of 0.49% of open access journals' articles but only 0.35% of closed access journals' articles, despite the higher impact factors.

Conclusion

This article examined in unprecedented detail and scale how the English language Wikipedia references the scientific literature. Of central interest was the relationship between an articles' academic status and accessibility on its probability of being referenced in Wikipedia. In the appendix, we make a cursory attempt to extend this analysis to the world's 50 largest Wikipedias. Previous studies have focused only on the role of academic status on referencing in the English Wikipedia and have often ignored unreferenced articles. In contrast, we began by identifying an enormous (~19.4MM articles) corpus of scientific literature that scientists routinely cite, found a subset of this literature for which Wikipedia editors demonstrate demand, and estimated a statistical model to identify the features of journals that predict referencing.

We found that a journal's academic status (impact factor) routinely predicts its appearance on Wikipedia. We also demonstrated, for the first time, that a journal's accessibility (open access policy) generally increases probability of referencing on Wikipedia as well, albeit less consistently than its impact factor. The odds that an open access journal is referenced on the English Wikipedia are about 47% higher compared to closed access, paywall journals. Moreover,

of closed access journals, those with high impact factors are also significantly more likely to appear in the English Wikipedia. Therefore, editors of the English Wikipedia act as "distillers" of high quality science by interpreting and distributing otherwise closed access knowledge to a broad public audience, free of charge. Moreover, the English Wikipedia, as a platform, acts as an "amplifier" for the (already freely available) open access literature by preferentially broadcasting its findings to millions.

Limitations and directions for future research

Our findings are not without limitations. First and foremost, it bears emphasis that this study did not investigate the nature of Wikipedia's sources as a whole (see Ford et al., 2013 for an excellent examination of sources). Only a fraction of Wikipedia's references use the scientific literature, and this is the subset on which we focused. Consequently the present study cannot address the concern expressed by others, e.g. (Luyt & Tan, 2010), that sources outside the scientific literature are used too heavily in scientific articles. Second, the study was cross-sectional in nature; it is conceivable that open access articles differ from closed access, paywall articles in their relevance to Wikipedia. Future work can test the potential confounding factor of unmeasured relevance by observing reference rates for articles which have been experimentally assigned to open and closed-access statuses, as has been done by some psychology journals (Davis et al., 2008).

Third, the study measured accessibility of articles by the open-access policy of the publishing journals. However, many articles in paywall journals are made freely available by their authors or third parties (Björk et al., 2014; Wren, 2005a). The resulting error in the measurement of accessibility may bias the observed advantage of open access in either direction: if open access *articles* from paywall *journals*, erroneously coded as closed access, are referenced

at higher rates than the journals' truly closed access articles (Gargouri et al., 2010; Harnad & Brody, 2004), the *true* advantage of open access will be even higher than we observed. In the (unlikely) case that open access articles in paywall journals are referenced less than closed access articles, the observed open access advantage will be an overestimate. The academic status of articles is also operationalized by a journal characteristic – its impact factor. In fact, many articles out- or under-perform their journal's impact factor. While this measurement error likely adds noise to the data, it probably does so without biasing the estimated effect of impact factor on referencing in one direction or another.

The impact of open access science

The chief finding of this study bears emphasis. We believe the existing discussion of open access has focused too narrowly on the academic literature. Early results showing that open access improves scientific outcomes such as citations have been tempered by newer experimental evidence showing small to null causal effects, and a lively debate has ensued. Our research shifts focus to diffusion, showing that open access policies have a tremendous impact on the diffusion of science to the broader general public through an intermediary like Wikipedia. This effect, previously a matter primarily of speculation, has empirical support. As millions of people search for scientific information on Wikipedia, the science they find distilled and referenced in its articles consists of a disproportionate quantity of open access science.

Acknowledgements

This research was enabled by grant 39147 to the Metaknowledge Network by the John Templeton Foundation. An earlier version of this work was presented at the Wikipedia Workshop of the 9th International Conference on Web and Social Media, Oxford, UK. We thank the reviewers for perceptive comments that greatly improved this article.

Appendix 5

Table A5.1. Most common sources referenced using the *cite journal* template that are not indexed by *Scopus*.

Journal name	Times referenced
BILLBOARD	1630
NATIONAL PARK SERVICE	539
ROYAL GAZETTE	523
BULL AMER MATH SOC	506
FLIGHT INTERNATIONAL	455
BAH NEWS	385
NEW YORK TIMES	369
ROLLING STONE	360
ENTERTAINMENT WEEKLY	342
J BOMBAY NAT HIST SOC	314
WHOS WHO	312
BIZJOURNALSCOM	288
THE GUARDIAN	287
THE LANCET	281
VARIETY MAGAZINE	270
INSIDE SOAP	239
TIME MAGAZINE	238
BIOCHIMICA ET BIOPHYSICA ACTA	201
EDMONTON JOURNAL	201
SPORTS ILLUSTRATED	197

Non-English Wikipedias

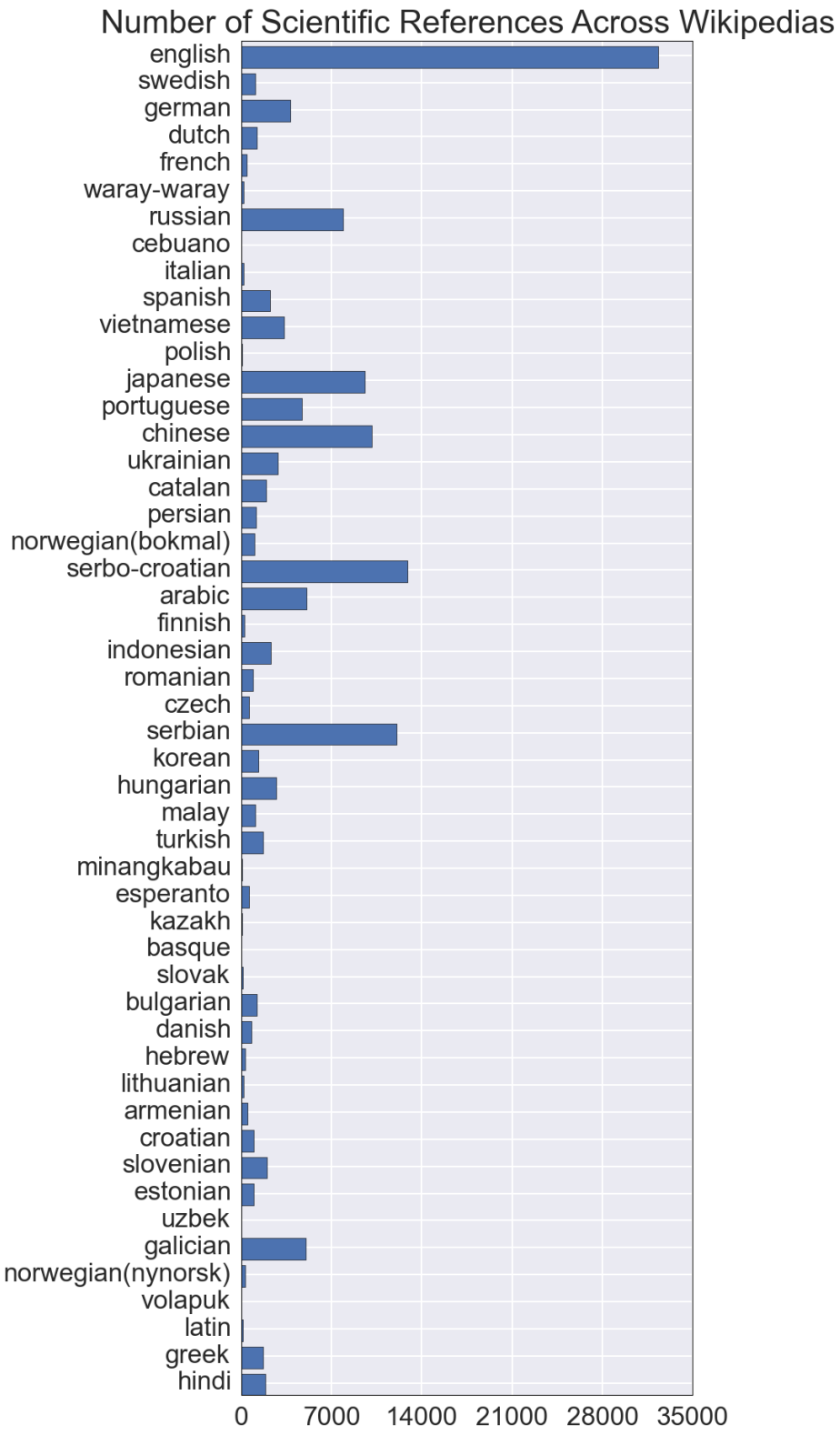
Non-English Wikipedias have been noticeably neglected by the research community (Mesgari, Okoli, Mehdi, Nielsen, & Lanamäki, 2015; Schroeder & Taylor, 2015). It is thus important to test whether any of the findings of this article extend to the millions of articles in non-English Wikipedias. Below we present an exploratory analysis of scientific references in the 49 largest non-English Wikipedias.

Data

Database dumps of the 49 largest non-English Wikipedias were downloaded on 2015-05-10. For each of these, we extracted tags containing "journal" or "doi". Thus the process for obtaining scientific references in non-English Wikipedias did not take into account language-specific tags. Non-English Wikipedias may also reference domestic scientific journals that are not indexed by *Scopus*. Thus, this exploratory approach surely undercounts scientific references to non-English Wikipedias.

The English Wikipedia referenced by far the greatest number of unique articles. Figure A5.1 displays the number of unique articles referenced in other Wikipedias, sorted by size (total articles).

Figure A5.1. Number of unique scientific articles referenced on the 50 largest Wikipedias.



Empirical Strategy

Certainly not all findings published in the academic literature belong on Wikipedia. Only small subsets of published findings are important and notable enough to be referenced in Wikipedia. Ideally, studies of how Wikipedia editors reference sources should explain which items in this smaller subset are and are not referenced. Nevertheless, previous studies have struggled to distinguish the candidate articles that are at risk for reference from those that do not belong on Wikipedia. Yet, to model referencing decisions with *all* articles – including the dozens of millions of articles never referenced on Wikipedia – is likely to result in a model that predicts that no article will ever be referenced. Consequently most studies have voluntarily hobbled themselves by simply modeling only on the subset of referenced articles.

Here we propose a compromise strategy based on "demonstrated demand." The idea is simple: articles are at risk for reference if *other* articles on the same topic are referenced. Topical reference indicates that there is demand from Wikipedia editors for literature on the topic and that an article's characteristics (e.g. accessibility, status) may determine which of the candidate articles an editor finds and references. Conversely, if articles on a given topic are never referenced, it is likely that Wikipedia editors do not "demand" literature on this topic, no matter the accessibility or status of the supply. Demonstrated demand exists at the level of topics and, like accessibility and status, we identify an article's topic at the journal level. Demonstrable demand is also a language-dependent metric: some Wikipedias may lack editors with expertise or interest in, for example, dentistry, thereby consigning all dentistry journals to irrelevance with regards to referencing decisions (but not irrelevant for analysis of coverage, of course). To calculate demonstrated demand we identify for each journal its topical "neighbors" and assign demand of 0 if none of these neighbors are ever referenced in a particular Wikipedia.

We calculate demonstrated demand for a journal through its topical neighbors, which are defined as other journals that share at least one narrow (ASJC) subject code. Only 1 journal, *Prevenzione & assistenza dentale*, had no neighbors while the mean neighborhood size was 144.8. Figure A5.2 displays the distribution (kde) of neighborhood size.

Figure A5.2. Distribution of the topical neighborhood sizes of journals. On average journals had 144.8 other journals that addressed the same topic(s).

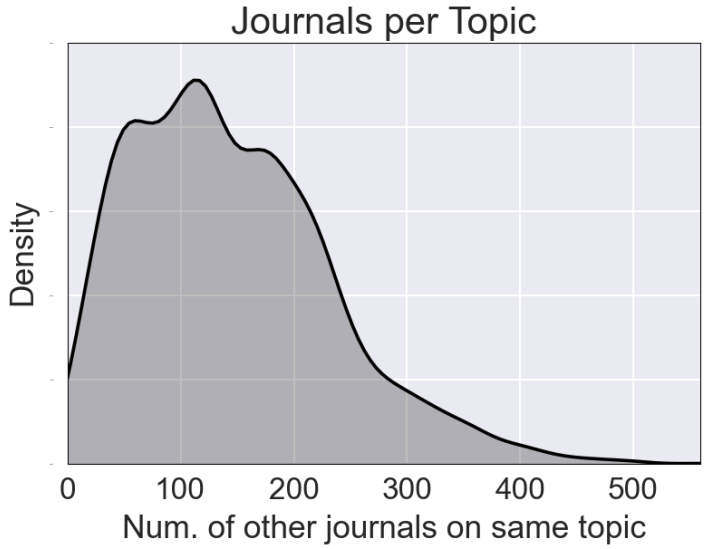


Table A5.2 contains the percentages of journals that were excluded from estimating models in each language. This percentage varies widely. For example, only 0.17% of journals were not used for the English Wikipedia model, 49.87% for Slovak, and a 100% for Volapuk. These numbers correspond directly to demonstrable demand for various research literatures by the editors of each Wikipedia. While the English Wikipedia references ~32,000 articles from top journals, the Slovak Wikipedia references only 108 and Volapuk references 0.

Table A5.2. Percent of journal data that is not used in estimates language-specific models (demonstrated demand = 0). These percentages indicate the portion of research areas for which there is no demonstrable demand from (language-specific) Wikipedia editors.

Wikipedia language	Percent excluded (weight=0)
chinese	1.28
russian	1.28
japanese	1.36
arabic	1.47
vietnamese	2.01
portuguese	2.16
german	2.30
spanish	2.62
indonesian	2.95
hindi	3.62
hungarian	3.73
ukrainian	4.17
slovenian	4.25
persian	4.40
serbian	4.92
greek	4.99
turkish	5.63
serbo-croatian	6.35
malay	6.85
bulgarian	7.12
dutch	7.31
catalan	7.88
danish	8.36
swedish	9.78
romanian	10.08
korean	10.08
estonian	11.73
galician	12.13
norwegian(bokmal)	12.36
czech	12.84
french	13.68
croatian	14.52
hebrew	20.93
armenian	22.62
waray-waray	25.32
esperanto	26.37
finnish	26.81
italian	29.16

Table A5.2 continued	
lithuanian	34.00
norwegian(nynorsk)	40.01
latin	43.67
slovak	49.87
polish	62.69
uzbek	78.91
minangkabau	80.88
kazakh	81.44
basque	86.30
cebuano	86.95
volapuk	99.98

Results

From each Wikipedia's model, two parameters are of focal interest: the odds ratio (of probability of referencing) when open access is True, and odds ratio when (log of) impact factor increases by 1 unit. Table A5.3 shows these odds ratios and associated p-values for each Wikipedia.

Table A5.3. Odds ratios and associated *p*-values for *open access* and (log) *impact factor* for 50 Wikipedias.

Wikipedia Language	**open access odds ratio**	**open access p-value**	**Ln(impact factor) odds ratio**	**Ln(impact factor) p-value**
arabic	0.923	0.258	2.189	0.000
armenian	1.052	0.802	2.669	0.000
basque	0.000	1.000	1.059	0.890
bulgarian	0.936	0.623	2.256	0.000
catalan	1.452	0.000	1.963	0.000
cebuano	0.000	1.000	1.989	0.093
chinese	1.337	0.000	2.257	0.000
croatian	0.651	0.009	2.230	0.000
czech	1.247	0.230	2.258	0.000
danish	0.722	0.120	2.190	0.000
dutch	1.743	0.000	2.238	0.000
english	1.471	0.000	1.878	0.000
esperanto	1.114	0.591	1.245	0.001
estonian	1.221	0.151	2.705	0.000
finnish	0.666	0.300	1.576	0.000

Table A5.3 continued				
french	0.850	0.550	2.030	0.000
galician	1.464	0.000	2.176	0.000
german	1.755	0.000	2.264	0.000
greek	0.798	0.078	2.008	0.000
hebrew	1.191	0.531	1.906	0.000
Hindi	0.757	0.029	2.113	0.000
hungarian	0.749	0.005	1.804	0.000
indonesian	0.886	0.242	2.467	0.000
italian	0.638	0.299	2.072	0.000
japanese	2.577	0.000	1.865	0.000
kazakh	0.000	1.000	1.759	0.114
korean	1.246	0.055	1.944	0.000
latin	0.812	0.637	1.975	0.000
lithuanian	1.035	0.923	2.345	0.000
malay	1.211	0.157	2.214	0.000
minangkabau	2.927	0.314	1.661	0.224
norwegian(bokmal)	0.866	0.437	2.328	0.000
norwegian(nynorsk)	0.588	0.109	1.510	0.000
persian	0.941	0.678	2.210	0.000
polish	0.588	0.480	2.330	0.000
portuguese	1.527	0.000	2.076	0.000
romanian	0.903	0.545	2.178	0.000
russian	1.419	0.000	2.086	0.000
serbian	3.824	0.000	1.516	0.000
serbo-croatian	3.761	0.000	1.518	0.000
slovak	1.943	0.157	3.249	0.000
slovenian	0.926	0.487	2.389	0.000
spanish	1.913	0.000	1.698	0.000
swedish	3.745	0.000	2.094	0.000
turkish	1.262	0.021	2.956	0.000
ukrainian	0.818	0.030	2.566	0.000
uzbek	0.000	1.000	2.963	0.012
vietnamese	0.966	0.682	2.143	0.000
volapuk	0.588	0.480	2.330	0.000
waray-waray	2.104	0.017	2.172	0.000

While earlier results showed that both accessibility and status increase the odds that a journal will be referenced in the English Wikipedia, the relative strength of these effects varies across languages. Some Wikipedias, like the Turkish, prioritize a journal's academic status over accessibility; the odds of referencing high status journals are nearly 200% higher than lower status journals. Other Wikipedias, like the Serbian, prioritize accessibility over status; the odds of referencing an open access journals are ~275% higher than a paywall journal.

Intuition and previous work suggests poorer countries rely on open access literature more (J. A. Evans & Reimer, 2009), yet this pattern is not apparent in Figure 8. For example, India and Ukraine, relatively poor countries naturally associated with the Hindi and Ukrainian Wikipedias, actually exhibit a small preference against open access literature, while a relatively wealthy country like Sweden has a Wikipedia that exhibits a huge preference for open access literature. The unexpected patterns may be due to the influence of bots (Steiner, 2014). For example, about a third of all articles on the Swedish Wikipedia were created by a bot (Jervell, 2014). Idiosyncrasies of the small number of human and non-human entities that edit science in non-English Wikipedias may thus play a larger role than gross cross-national patterns.

It bears emphasis that this analysis of references in non-English Wikipedias is exploratory. Further work should extract references in a way that is sensitive to each Wikipedia's language and conventions. Such analysis may reveal differences in how scientific content found in Wikipedia across languages is differentially embedded in or husbanded by local scientific communities.

CHAPTER 6. CONCLUSION

This dissertation presented four case studies of valuation practices of scientific manuscripts and the consequences of these practices, particularly for the field of sociology. Each case focused on a different step of the evaluation process, as illustrated by Figure 1.1 of the Introduction. These foci included the revision process and how the relative values it revealed may affect the investments authors make into their data analyses and theoretical frames; reviewers' reports and decisions and the citing decisions of the academic audience; the editorial bias for positive findings and how it affects the robustness of the literature that uses the General Social Survey; and the citing decisions of a non-academic audience of Wikipedians. This chapter attempts to synthesize these various case studies and make explicit the contributions they make to a general theory of scientific valuation. Many of these conclusions are not new; similar arguments may be found in disparate areas of the tremendously voluminous literature on peer review. One of the contributions, then, is to collect these arguments together. Moreover, many of the conclusions are speculative. The main value of the latter is to identify directions for future work.

The first set of remarks concerns the conceptual shift away from disagreement as normatively problematic. I argue that conceptualization of evaluation as deploying rigid evaluative criteria has resulted in an unproductive preoccupation in the literature with the large amount of disagreement between reviewers. A conceptualization of review as deploying flexible values, some of which may be in opposition, reduces the normative significance of disagreement and shifts attention to its social structure. The value of "interestingness" is especially crucial in review and underlies two of the dissertation's case studies.

The second set of remarks concerns the shift away from explaining individual decisions to the consequences of prioritizing particular values during review. The linking of variation in review practices with consequences, I argue, is a relatively fertile area and one that is the likely source of reliable policy recommendations.

From disagreement to the social structure of disagreement

As mentioned in the Introduction, the literature on peer review is now well over 3,000 articles and books. Many if not most of these studies measure disagreement between reviewers, which is found to be consistently high (Bornmann, 2011a; Cicchetti, 1991b). The preoccupation with the *amount* of disagreement derives in part from its practical consequences: if favorable review outcomes depend on the fortunate selection of a favorable review panel, that is they are unpredictable by the grant applicant or author, then the applicants and authors will rationally underinvest into the quality of their submissions. A second cause of the preoccupation is the view, dating back to Zuckerman and Merton's seminal 1971 article (Zuckerman & Merton, 1971), that review decisions should be motivated by the desirable "universalistic" criteria, but they may be corrupted by the undesirable "particularistic" concerns. These particularistic concerns corrupt not only the validity of the review judgments but their reliability as well: if scientists disagree about the value of a particular application or manuscript, they must disagree about their opinion of the author or similar particularistic factors.

After more than 40 years, it is difficult to identify concrete insights or policy implications produced by this stream of research. To reveal which motivations lead to this or that review outcome has proven extremely difficult from observational data, and the scientific community has fiercely resisted peer review experiments (Chubin & Hackett, 1990). The proxies of authors' "particularistic" characteristics, such as their academic rank, suffer from obvious confounding

with “universalistic” characteristics, such as academic achievement (Lee et al., 2013). Moreover, even these proxies are associated with review outcomes only weakly and inconsistently (V. Bakanic et al., 1987; Chubin & Hackett, 1990; Lee et al., 2013). Perhaps more importantly, recent studies by sociologists of culture and others have contested the very division of motivations into “universalistic” and “particularistic.” Scientists, the argument goes, can and do disagree for a variety of normatively neutral reasons (Boudreau, Guinan, Lakhani, & Riedl, 2016; J. Guetzkow, Lamont, & Mallard, 2004; Michèle Lamont, 2009b; Mallard et al., 2009; Travis & Collins, 1991).

The normatively neutral causes of disagreement were discussed as early as 1973 by Thomas Kuhn (Kuhn, 1977a, Chapter 13). During theory choice, Kuhn wrote, “two men fully committed to the same list of criteria for choice may nevertheless reach different conclusions. Perhaps they interpret ‘simplicity’ differently or have different convictions about the range of fields within which the consistency criterion must be met. Or perhaps they agree about these matters but differ about the relative weights to be accorded to these or to other criteria” (Kuhn, 1977b, p. 334). Evaluative criteria are thus open to interpretation, but this does not imply that each scientist uses a completely idiosyncratic decision algorithm. What scientists do deploy collectively is better described as shared evaluative values, and the limited evidence indicates that scientists do tend to agree on what these values are (Chase, 1970; Kuhn, 1977a, Chapter 13; Michèle Lamont, 2009b).

The important, and largely open, question is not whether scientists disagree, but whether disagreement is socially structured, and what consequences this structure may have for the knowledge claims that pass through review. Existing work has identified large differences in cognitive commitments *across* disciplines (J. Guetzkow et al., 2004; Michèle Lamont, 2009b),

but the structure of *intra*-disciplinary cognitive commitments (Boudreau et al., 2016; Travis & Collins, 1991), and especially their consequences (Björk & Solomon, 2012; Kuhn, 1977a, Chapter 13), are fertile for exploration.

The value of "interestingness" and its consequences

A focus on consequences also serves as an important bridge between the qualitative work that reveals interpretive flexibility entailed in evaluation and the earlier, normatively oriented work. It is unclear whether it is possible at present to agree on whether particular considerations during review, such as an author's identity, are normatively problematic, neutral, or even beneficial. It is much more likely that consensus may be reached regarding the characteristics of the literature scholars want evaluation processes to promote. For instance, scholars of all persuasions are likely to value a literature that consists of robust, reproducible findings. Many scholars are also likely to value a risk-taking literature that pushes the boundaries of knowledge. It is important to understand how particular review practices promote or retard particular consequences (Boudreau et al., 2016; Huutoniemi, 2012; Langfeldt, 2006).

How interesting or important a piece of scholarship is, regardless of its technical merit, is often a crucial consideration in review. Where in the review process this evaluation is made, if it is made at all, varies widely (Chubin & Hackett, 1990). As bibliometric impact becomes the conventional metric by which to evaluate academic work, it is important to identify which step of the review process is responsible for ensuring that favorably reviewed research produces impact. The fact that this responsibility is distinct from ensuring technical merit alone, and that it may be situated in various steps of review, is the contribution of Chapter 3 of this dissertation. The chapter used the review files of the *American Sociological Review* to show that reviewers in the late 70s and early 80s did not consider assessment of potential impact part of their mandate. The

finding that review scores are not associated with citation impact should not be interpreted as a failure of the review system as a whole, as is commonly concluded.

Chapter 4 also hinges on the value of interestingness. The bias for positive findings, which was here taken for granted rather than measured, stems from a desire by editors to publish the presumably relatively rare claims about what does "work," rather than the plentiful claims about what does not. A consequence of this bias is to motivate authors to produce positive findings, possibly at the expensive of robustness (Open Science Collaboration, 2015). Chapter 4 shows that the effects of this bias in quantitative sociology are relatively minor. Claims that use the General Social Survey are relatively reproducible soon after publication. However, they become more fragile with time, and this effect overshadows authors' reactions to editors' bias toward positive findings. The finding contrasts starkly with those emerging from social psychology; I suspect the relatively high statistical power of the General Social Survey is critical in minimizing the effects of editorial bias.

Undertheorized consequences

Some consequences engendered by particular review practices have received consistent attention. The foremost of these is the extent to which review promotes or discourages risky research (Boudreau et al., 2016; Chubin & Hackett, 1990; Langfeldt, 2006; Lee et al., 2013; Roy, 1985). Yet review practices are likely to have a number of other consequences as well. Chapter 2 focuses on the character of revisions to submitted sociological manuscripts that authors, reviewers, and editors negotiate to identify the relative value placed upon data and analysis on one hand and theoretical framing on the other. Data and analysis appear to be relatively "expensive," while theoretical frames are relatively cheap and plentiful. If peer review is a context in which a discipline instantiates its epistemic values, authors will react by investing

more heavily into aspects of research with the highest value – data and analysis. This appears to be the case in contemporary sociology. What the discipline's literature would look like if peer review placed the higher "price" on theoretical framing remains open to speculation. Nevertheless, the chapter shows it is important to broaden the list of potential consequences of review practices.

Academic peer review decisions also have consequences in spheres outside of academic science. The interaction between peer review and these spheres have received very limited attention (for an exception see Jasanoff, 1985). Chapter 5 focuses on one such sphere – Wikipedia, the world's largest encyclopedia. Wikipedia is a major importer of the scientific literature, using hundreds of thousands of articles as source material. The chapter shows that Wikipedians import the same status hierarchy of academic journals that is used by scientists. An implication is that the values that dictate the review decisions of editorial boards oriented to scientific audiences, including the value of interestingness and the bias for positive findings, can unexpectedly "leak" into seemingly unrelated spheres. This pathway between the scientific community and the lay public, and how review practices affect it, is deserving of future attention.

REFERENCES

Abbott, A. (1988). Transcending General Linear Reality. *Sociological Theory*, *6*(2), 169–186. http://doi.org/10.2307/202114

Abbott, A. (1997). Seven Types of Ambiguity. *Theory and Society*, *26*(2/3), 357–391.

Abbott, A. (2014). *Digital Paper: A Manual for Research and Writing with Library and Internet Materials*. Chicago ; London: University Of Chicago Press.

Abbott, A., & Alexander, J. C. (2004). *Methods of Discovery: Heuristics for the Social Sciences*. New York: W. W. Norton & Company.

Abend, G. (2006). Styles of Sociological Thought: Sociologies, Epistemologies, and the Mexican and U.S. Quests for Truth*. *Sociological Theory*, *24*(1), 1–41. http://doi.org/10.1111/j.0735-2751.2006.00262.x

Abend, G. (2008). The Meaning of "Theory"*. *Sociological Theory*, *26*(2), 173–199. http://doi.org/10.1111/j.1467-9558.2008.00324.x

Alan G. Gross. (1990). *The Rhetoric of Science*. Cambridge, Mass: Harvard University Press.

ALM, P., Fenner, M., & Lin, J. (2014, June 7). An analysis of Wikipedia references across PLOS publications. Retrieved from http://figshare.com/articles/An_analysis_of_Wikipedia_references_across_PLOS_publications/1048991

altmetrics: a manifesto – altmetrics.org. (n.d.). Retrieved from http://altmetrics.org/manifesto/

ATLAS Collaboration. (2012). Combined search for the Standard Model Higgs boson in pp collisions at sqrt(s) = 7 TeV with the ATLAS detector. *Physical Review D*, *86*(3). http://doi.org/10.1103/PhysRevD.86.032003

Babbie, E. R. (2014). *The practice of social research*.

Bail, C. A. (2008). The Configuration of Symbolic Boundaries against Immigrants in Europe. *American Sociological Review*, *73*(1), 37–59. http://doi.org/10.1177/000312240807300103

Bakanic, E. Y. (1986). *Tracing Social Science: The Manuscript Review Process* (Ph.D.). University of Illinois at Urbana-Champaign, United States -- Illinois. Retrieved from http://search.proquest.com.proxy.uchicago.edu/pqdtglobal/docview/303417130/abstract/44446EC919EF45BCPQ/1

Bakanic, V., McPhail, C., & Simon, R. J. (1987). The Manuscript Review and Decision-Making Process. *American Sociological Review*, *52*(5), 631–642. http://doi.org/10.2307/2095599

Bakanic, V., McPhail, C., & Simon, R. J. (1989). Mixed Messages: Referees' Comments on the Manuscripts They Review. *The Sociological Quarterly*, *30*(4), 639–654.

Baltay, C., & Rosenfeld, A. H. (1968). *Meson spectroscopy; a collection of articles.* New York: W. A. Benjamin. Retrieved from http://pi.lib.uchicago.edu/1001/cat/bib/1082525

Barber, R. F., & Candes, E. (2014). Controlling the False Discovery Rate via Knockoffs. *arXiv:1404.5609 [math, Stat]*. Retrieved from http://arxiv.org/abs/1404.5609

Bazerman, C. (1988). *Shaping Written Knowledge: The Genre and Activity of the Experimental Article in Science* (1ST edition). Madison, Wis: Univ of Wisconsin Pr.

Beirne, P. (1987). Adolphe Quetelet and the Origins of Positivist Criminology. *American Journal of Sociology*, *92*(5), 1140–1169.

Benjamini, Y., & Hochberg, Y. (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.

Björk, B.-C., Laakso, M., Welling, P., & Paetau, P. (2014). Anatomy of green open access. *Journal of the Association for Information Science and Technology*, *65*(2), 237–250. http://doi.org/10.1002/asi.22963

Björk, B.-C., & Solomon, D. (2012). Open access versus subscription journals: a comparison of scientific impact. *BMC Medicine*, *10*(1), 73. http://doi.org/10.1186/1741-7015-10-73

Boas, F. (1894). Human Faculty as Determined by Race. In G. W. Stocking (Ed.), *The shaping of American anthropology, 1883-1911; a Franz Boas reader.* (pp. 221–242). New York: Basic Books.

Boas, F. (1897). *Race, language and culture*. New York: The Macmillan Co.

Bonferroni, C. E. (1936). *Teoria statistica delle classi e calcolo delle probabilità*. Firenze: Seeber.

Bornmann, L. (2011a). Scientific peer review. *Annual Review of Information Science and Technology*, *45*(1), 197–245. http://doi.org/10.1002/aris.2011.1440450112

Bornmann, L. (2011b). Scientific peer review. *Annual Review of Information Science and Technology*, *45*(1), 197–245. http://doi.org/10.1002/aris.2011.1440450112

Bornmann, L., & Daniel, H. (2008a). What do citation counts measure? A review of studies on citing behavior. *Journal of Documentation*, *64*(1), 45–80. http://doi.org/10.1108/00220410810844150

Bornmann, L., & Daniel, H.-D. (2008b). Selecting manuscripts for a high-impact journal through peer review: A citation analysis of communications that were accepted by Angewandte Chemie International Edition, or rejected but published elsewhere. *Journal of the*

American Society for Information Science and Technology, *59*(11), 1841–1852. http://doi.org/10.1002/asi.20901

Bornmann, L., & Marx, W. (2014). The wisdom of citing scientists. *Journal of the Association for Information Science and Technology*, *65*(6), 1288–1292. http://doi.org/10.1002/asi.23100

Bornmann, L., Mutz, R., & Daniel, H.-D. (2010). A Reliability-Generalization Study of Journal Peer Reviews: A Multilevel Meta-Analysis of Inter-Rater Reliability and Its Determinants. *PLoS ONE*, *5*(12), e14331. http://doi.org/10.1371/journal.pone.0014331

Bornmann, L., Nast, I., & Daniel, H.-D. (2008). Do editors and referees look for signs of scientific misconduct when reviewing manuscripts? A quantitative content analysis of studies that examined review criteria and reasons for accepting and rejecting manuscripts for publication. *Scientometrics*, *77*(3), 415–432. http://doi.org/10.1007/s11192-007-1950-2

Bornmann, L., Schier, H., Marx, W., & Daniel, H.-D. (2012). What factors determine citation counts of publications in chemistry besides their quality? *Journal of Informetrics*, *6*(1), 11–18. http://doi.org/10.1016/j.joi.2011.08.004

Bornmann, L., Thor, A., Marx, W., & Schier, H. (Forthcoming). The application of bibliometrics to research evaluation in the humanities and social sciences: an exploratory study using normalized Google Scholar data for the publications of a research institute. *Journal of the Association for Information Science and Technology*.

Boudreau, K. J., Guinan, E. C., Lakhani, K. R., & Riedl, C. (2016). Looking Across and Looking Beyond the Knowledge Frontier: Intellectual Distance, Novelty, and Resource Allocation in Science. *Management Science*.

Bourdieu, P. (1975). The specificity of the scientific field and the social conditions of the progress of reason. *Social Science Information*, *14*(6), 19–47. http://doi.org/10.1177/053901847501400602

Broadus, R. N. (1983). An investigation of the validity of bibliographic citations. *Journal of the American Society for Information Science*, *34*(2), 132–135. http://doi.org/10.1002/asi.4630340206

Brunsma, D., Prasad, M., & Zuckerman, E. (2013). Strategies for Reviewing Manuscripts. Retrieved from http://www.asanet.org/documents/asa/pdfs/Review_Times_in_Sociology.pdf

Butler, B., Joyce, E., & Pike, J. (2008). Don'T Look Now, but We've Created a Bureaucracy: The Nature and Roles of Policies and Rules in Wikipedia. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 1101–1110). New York, NY, USA: ACM. http://doi.org/10.1145/1357054.1357227

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munafò, M. R. (2013). Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, *14*(5), 365–376. http://doi.org/10.1038/nrn3475

Calcagno, V., Demoinet, E., Gollner, K., Guidi, L., Ruths, D., & Mazancourt, C. de. (2012). Flows of Research Manuscripts Among Scientific Journals Reveal Hidden Submission Patterns. *Science*, 1227833. http://doi.org/10.1126/science.1227833

Camic, C. (1995). Three Departments in Search of a Discipline: Localism and Interdisciplinary Interaction in American Sociology, 1890—1940. *Social Research*, *62*(4), 1003–1033.

Camic, C., Gross, N., & Lamont, M. (Eds.). (2011). *Social Knowledge in the Making* (1 edition). Chicago : London: University Of Chicago Press.

Camic, C., & Xie, Y. (1994). The Statistical Turn in American Social Science: Columbia University, 1890 to 1915. *American Sociological Review*, *59*(5), 773–805. http://doi.org/10.2307/2096447

Campanario, J. M. (1998a). Peer Review for Journals as it Stands Today—Part 1. *Science Communication*, *19*(3), 181–211. http://doi.org/10.1177/1075547098019003002

Campanario, J. M. (1998b). Peer Review for Journals as it Stands Today—Part 2. *Science Communication*, *19*(4), 277–306. http://doi.org/10.1177/1075547098019004002

Chase, J. M. (1970). Normative Criteria for Scientific Publication. *The American Sociologist*, *5*(3), 262–265.

Christ, C. F. (1985). Early Progress in Estimating Quantitative Economic Relationships in America. *The American Economic Review*, *75*(6), 39–52.

Chubin, D. E., & Hackett, E. J. (1990). *Peerless Science: Peer Review and U.S. Science Policy*. Albany, N.Y: State Univ of New York Pr.

Cicchetti, D. V. (1991a). The reliability of peer review for manuscript and grant submissions: A cross-disciplinary investigation. *Behavioral and Brain Sciences*, *14*(01), 119–135. http://doi.org/10.1017/S0140525X00065675

Cicchetti, D. V. (1991b). The reliability of peer review for manuscript and grant submissions: A cross-disciplinary investigation. *Behavioral and Brain Sciences*, *14*(01), 119–135. http://doi.org/10.1017/S0140525X00065675

Clemens, E. S., Powell, W. W., McIlwaine, K., & Okamoto, D. (1995). Careers in Print: Books, Journals, and Scholarly Reputations. *American Journal of Sociology*, *101*(2), 433–494.

Cohen, L. E., & Felson, M. (1979). Social Change and Crime Rate Trends: A Routine Activity Approach. *American Sociological Review*, *44*(4), 588–608. http://doi.org/10.2307/2094589

Coleman, J. S., Campbell, E. Q., Hobson, C. J., McPartland, J., Mood, A. M., Weinfield, F. D., & York, R. L. (1966). *Equality of Educational Opportunity Study* (No. OE-38001) (p. 749). U.S. Department of Health, Education and Welfare.

Cole, S. (Ed.). (2001). *What's Wrong with Sociology?*. New Brunswick, N.J: Transaction Publishers.

Cole, S., Cole, J. R., & Simon, G. A. (1981). Chance and Consensus in Peer Review. *Science*, *214*(4523), 881–886. http://doi.org/10.2307/1686309

Comte, A. (1830). *Cours de philosophie positive. [Tome 4] / par M. Auguste Comte,...* Rouen frères (Bachelier) (Paris). Retrieved from http://gallica.bnf.fr/ark:/12148/bpt6k76270k

Danthi, N., Wu, C. O., Shi, P., & Lauer, M. S. (2014). Percentile Ranking and Citation Impact of a Large Cohort of NHLBI-Funded Cardiovascular R01 Grants. *Circulation Research*, CIRCRESAHA.113.302656. http://doi.org/10.1161/CIRCRESAHA.114.302656

Davis, P. M. (2010). Does open access lead to increased readership and citations? A randomized controlled trial of articles published in APS journals. *The Physiologist*, *53*(6), 197, 200–201.

Davis, P. M. (2011). Open access, readership, citations: a randomized controlled trial of scientific journal publishing. *The FASEB Journal*, *25*(7), 2129–2134. http://doi.org/10.1096/fj.11-183988

Davis, P. M., Lewenstein, B. V., Simon, D. H., Booth, J. G., & Connolly, M. J. L. (2008). Open access publishing, article downloads, and citations: randomised controlled trial. *BMJ*, *337*, a568. http://doi.org/10.1136/bmj.a568

Davis, P. M., & Walters, W. H. (2011). The impact of free access to the scientific literature: a review of recent research. *Journal of the Medical Library Association : JMLA*, *99*(3), 208–217. http://doi.org/10.3163/1536-5050.99.3.008

Dear, P. (1985). Totius in Verba: Rhetoric and Authority in the Early Royal Society. *Isis*, *76*(2), 145–161.

Dudoit, S., & Van der Laan, M. J. (2008). *Multiple Testing Procedures with Applications to Genomics*. Springer. Retrieved from http://www.springer.com/life+sciences/biochemistry+%26+biophysics/book/978-0-387-49316-9

Duncan, O. D. (1984). *Notes on Social Measurement: Historical and Critical*. Russell Sage Foundation.

Dunnett, C. W., & Tamhane, A. C. (1992). A Step-Up Multiple Test Procedure. *Journal of the American Statistical Association*, *87*(417), 162–170. http://doi.org/10.2307/2290465

Dunn, O. J. (1959). Estimation of the Medians for Dependent Variables. *The Annals of Mathematical Statistics*, *30*(1), 192–197. http://doi.org/10.1214/aoms/1177706374

Dunn, O. J. (1961). Multiple Comparisons among Means. *Journal of the American Statistical Association*, *56*(293), 52–64. http://doi.org/10.1080/01621459.1961.10482090

England, P., Hermsen, J. M., & Cotter, D. A. (2000). The Devaluation of Women's Work: A Comment on Tam. *American Journal of Sociology*, *105*(6), 1741–1751.

Espeland, W. N., & Sauder, M. (2007). Rankings and reactivity: How public measures recreate social worlds. *American Journal of Sociology*, *113*(1), 1–40. http://doi.org/10.1086/517897

Espeland, W. N., & Stevens, M. L. (2008). A Sociology of Quantification. *European Journal of Sociology / Archives Européennes de Sociologie*, *49*(03), 401–436. http://doi.org/10.1017/S0003975609000150

Evans, J. A., & Reimer, J. (2009). Open Access and Global Participation in Science. *Science*, *323*, 1025.

Evans, P., & Krauthammer, M. (2011). Exploring the Use of Social Media to Measure Journal Article Impact. *AMIA Annual Symposium Proceedings*, *2011*, 374–381.

Eysenbach, G. (2006a). Citation Advantage of Open Access Articles. *PLoS Biol*, *4*(5), e157. http://doi.org/10.1371/journal.pbio.0040157

Eysenbach, G. (2006b). The Open Access Advantage. *Journal of Medical Internet Research*, *8*(2), e8. http://doi.org/10.2196/jmir.8.2.e8

Fleiss, J. L. (1981). *Statistical methods for rates and proportions* (Second edition). New York: John Wiley.

Ford, H., Sen, S., Musicant, D. R., & Miller, N. (2013). Getting to the Source: Where Does Wikipedia Get Its Information from? In *Proceedings of the 9th International Symposium on Open Collaboration* (pp. 9:1–9:10). New York, NY, USA: ACM. http://doi.org/10.1145/2491055.2491064

Fox, M. F. (1989). Disciplinary fragmentation, peer review, and the publication process. *The American Sociologist*, *20*(2), 188–191. http://doi.org/10.1007/BF02691858

Fox, M. F., & Firebaugh, G. (1992). Confidence in science: The gender gap. *Social Science Quarterly*, *73*(1), 101–113.

Franco, A., Malhotra, N., & Simonovits, G. (2014). Publication bias in the social sciences: Unlocking the file drawer. *Science*, *345*(6203), 1502–1505. http://doi.org/10.1126/science.1255484

Freese, J. (2007). Replication Standards for Quantitative Social Science Why Not Sociology? *Sociological Methods & Research*, *36*(2), 153–172. http://doi.org/10.1177/0049124107306659

Gargouri, Y., Hajjem, C., Larivière, V., Gingras, Y., Carr, L., Brody, T., & Harnad, S. (2010). Self-Selected or Mandated, Open Access Increases Citation Impact for Higher Quality Research. *PLoS ONE*, *5*(10), e13636. http://doi.org/10.1371/journal.pone.0013636

Gaulé, P., & Maystre, N. (2011). Getting cited: Does open access help? *Research Policy*, *40*(10), 1332–1338. http://doi.org/10.1016/j.respol.2011.05.025

Gerber, A., & Malhotra, N. (2008). Do Statistical Reporting Standards Affect What Is Published? Publication Bias in Two Leading Political Science Journals. *Quarterly Journal of Political Science*, *3*(3), 313–326. http://doi.org/10.1561/100.00008024

Gerber, A. S., & Malhotra, N. (2008). Publication Bias in Empirical Sociological Research: Do Arbitrary Significance Levels Distort Published Research. *Sociological Methods & Research*, *37*(1), 3–30.

Gibson, L. (2013). Growing numbers. *University of Chicago Magazine*, *Sept-Oct 2013*. Retrieved from http://mag.uchicago.edu/law-policy-society/growing-numbers?msource=MAG10

Giddens, F. (1899). Exact Methods in Sociology. *Popular Science Monthly*, *56*, 145–59.

Giddings, F. H. (1901). *Inductive sociology; a syllabus of methods, analyses and classifications, and provisionally formulated laws.* New York; London: The Macmillan company; Macmillan & Co., Ltd.

Giddings, F. H. (1910). The Social Marking System. *American Journal of Sociology*, *15*(6), 721–740.

Gigerenzer, G. (1987). The Probabilistic Revolution in Psychology--An Overview. In *The probabilistic revolution, Vol. 1: Ideas in history; Vol. 2: Ideas in the sciences* (pp. 7–9). Cambridge, MA, US: The MIT Press.

Gilbert, G. N. (1976). The Transformation of Research Findings into Scientific Knowledge. *Social Studies of Science*, *6*(3/4), 281–306.

Gilbert, G. N. (1977). Referencing as Persuasion. *Social Studies of Science*, *7*(1), 113–122.

Gilbert, G. N., & Mulkay, M. (1984). *Opening Pandora's Box: A Sociological Analysis of Scientists' Discourse*. Cambridge Cambridgeshire ; New York: Cambridge University Press.

Glaser, B., & Strauss, A. (1999). *The Discovery of Grounded Theory: Strategies for Qualitative Research*. Chicago: Aldine Transaction.

González-Pereira, B., Guerrero-Bote, V. P., & Moya-Anegón, F. (2010). A new approach to the metric of journals' scientific prestige: The SJR indicator. *Journal of Informetrics*, *4*(3), 379–391. http://doi.org/10.1016/j.joi.2010.03.002

Goodman, S. N., Berlin, J., Fletcher, S. W., & Fletcher, R. H. (1994). Manuscript Quality before and after Peer Review and Editing at Annals of Internal Medicine. *Annals of Internal Medicine*, *121*(1), 11–21. http://doi.org/10.7326/0003-4819-121-1-199407010-00003

Gross, A. G. (1990). Persuasion and peer review in science: Habermas's ideal speech situation applied. *History of the Human Sciences*, *3*(2), 195–209. http://doi.org/10.1177/095269519000300203

Guetzkow, J., Lamont, M., & Mallard, G. (2004). What Is Originality in the Humanities and the Social Sciences? *American Sociological Review*, *69*(2), 190–212.

Guetzkow, J., Lamont, M., & Mallard, G. (2004). What is Originality in the Humanities and the Social Sciences? *American Sociological Review*, *69*(2), 190–212.

Gusfield, J. (1976). The Literary Rhetoric of Science: Comedy and Pathos in Drinking Driver Research. *American Sociological Review*, *41*(1), 16–34. http://doi.org/10.2307/2094370

Gwet, K. L. (2012). *Handbook of Inter-Rater Reliability (3rd Edition): The Definitive Guide to Measuring the Extent of Agreement Among Multiple Raters.* Advanced Analytics Press.

Hardin, J. W., & Hilbe, J. M. (2012). *Generalized Linear Models and Extensions, Third Edition* (3 edition). College Station, Tex: Stata Press.

Harnad, S. (2000). The invisible hand of peer review [Journal (On-line/Unpaginated)]. Retrieved September 8, 2014, from http://cogprints.org/1646/

Harnad, S., & Brody, T. (2004). Comparing the Impact of Open Access (OA) vs. Non-OA Articles in the Same Journals. *D-Lib Magazine*, *10*(6). Retrieved from http://eprints.soton.ac.uk/260207/

Heilman, J. M., & West, A. G. (2015). Wikipedia and Medicine: Quantifying Readership, Editors, and the Significance of Natural Language. *Journal of Medical Internet Research*, *17*(3), e62. http://doi.org/10.2196/jmir.4069

Hilts, V. L. (1973). Statistics and Social Science. In R. N. Giere & R. S. Westfall (Eds.), *Foundations of scientific method: the nineteenth century.* Bloomington: Indiana University Press.

Hilts, V. L. (1981). *Statist and statistician*. New York: Arno Press.

Hirschauer, S. (2010). Editorial Judgments: A Praxeology of "Voting" in Peer Review. *Social Studies of Science*, *40*(1), 71–103.

Holm, S. (1979). A Simple Sequentially Rejective Multiple Test Procedure. *Scandinavian Journal of Statistics*, *6*(2), 65–70.

Hubbard, R. (2004). Alphabet Soup Blurring the Distinctions Betweenp's anda's in Psychological Research. *Theory & Psychology*, *14*(3), 295–327. http://doi.org/10.1177/0959354304043638

Huutoniemi, K. (2012). Communicating and Compromising on Disciplinary Expertise in the Peer Review of Research Proposals. *Social Studies of Science*, 0306312712458478. http://doi.org/10.1177/0306312712458478

Ioannidis, J., & Doucouliagos, C. (2013). What's to Know About the Credibility of Empirical Economics? *Journal of Economic Surveys*, *27*(5), 997–1004. http://doi.org/10.1111/joes.12032

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

Ioannidis, J. P. A., Ntzani, E. E., Trikalinos, T. A., & Contopoulos-Ioannidis, D. G. (2001). Replication validity of genetic association studies. *Nature Genetics*, *29*(3), 306–309. http://doi.org/10.1038/ng749

Jamali, H. R., & Nikzad, M. (2011). Article Title Type and Its Relation with the Number of Downloads and Citations. *Scientometrics*, *88*(2), 653–661. http://doi.org/10.1007/s11192-011-0412-z

Jasanoff, S. (1985). Peer Review in the Regulatory Process. *Science, Technology, & Human Values*, *10*(3), 20–32.

Jeřábek, H. (2001). Paul Lazarsfeld—The Founder of Modern Empirical Sociology: A Research Biography. *International Journal of Public Opinion Research*, *13*(3), 229–244. http://doi.org/10.1093/ijpor/13.3.229

Jervell, E. E. (2014, July 14). For This Author, 10,000 Wikipedia Articles Is a Good Day's Work. *Wall Street Journal*. Retrieved from http://www.wsj.com/articles/for-this-author-10-000-wikipedia-articles-is-a-good-days-work-1405305001

Joe Wass. (2015, March 3). Real-time Stream of DOIs being cited in Wikipedia. Retrieved October 9, 2015, from http://crosstech.crossref.org/2015/03/real-time-stream-of-dois-being-cited-in-the-wikipedia.html

King, G., Tomz, M., & Wittenberg, J. (2000). Making the Most of Statistical Analyses: Improving Interpretation and Presentation. *American Journal of Political Science*, *44*(2), 347–361. http://doi.org/10.2307/2669316

Knorr-Cetina, K. (1981). *The manufacture of knowledge : an essay on the constructivist and contextual nature of science*. Oxford ; New York: Pergamon Press.

Knorr-Cetina, K. (1999). *Epistemic cultures: how the sciences make knowledge*. Cambridge, Mass.: Harvard University Press.

Knorr, K. D. (1977). Producing and reproducing knowledge: Descriptive or constructive? Toward a model of research production. *Social Science Information*, *16*(6), 669–696. http://doi.org/10.1177/053901847701600602

Knorr, K. D. (1979). Tinkering toward Success: Prelude to a Theory of Scientific Practice. *Theory and Society*, *8*(3), 347–376.

knorr, K. D., & knorr, D. (1978). on the relationship between laboratory research and published paper in science. Retrieved from https://www.ihs.ac.at/publications/ihsfo/fo132.pdf

Kuhn, T. S. (1977a). *The essential tension : selected studies in scientific tradition and change*. Chicago: University of Chicago Press.

Kuhn, T. S. (1977b). The Essential Tension: Tradition and Innovation in Scientific Research. In *The essential tension : selected studies in scientific tradition and change* (pp. 225–239). Chicago: University of Chicago Press.

Lakhani, K. R., & von Hippel, E. (2003). How open source software works: "free" user-to-user assistance. *Research Policy*, *32*(6), 923–943. http://doi.org/10.1016/S0048-7333(02)00095-1

Lakhani, K., & Wolf, R. G. (2003). *Why Hackers Do What They Do: Understanding Motivation and Effort in Free/Open Source Software Projects* (SSRN Scholarly Paper No. ID 443040). Rochester, NY: Social Science Research Network. Retrieved from http://papers.ssrn.com/abstract=443040

Lamb, E. (2012, July 17). 5 Sigma—What's That? | Observations, Scientific American Blog Network. Retrieved January 7, 2015, from http://blogs.scientificamerican.com/observations/2012/07/17/five-sigmawhats-that/

Lamont, M. (1987). How to Become a Dominant French Philosopher: The Case of Jacques Derrida. *American Journal of Sociology*, *93*(3), 584–622.

Lamont, M. (2009a). *How professors think : inside the curious world of academic judgment*. Cambridge, Mass.: Harvard University Press.

Lamont, M. (2009b). *How professors think : inside the curious world of academic judgment*. Cambridge, Mass.: Harvard University Press.

Lamont, M. (2012). Toward a Comparative Sociology of Valuation and Evaluation. *Annual Review of Sociology*, *38*(1), 201–221. http://doi.org/10.1146/annurev-soc-070308-120022

Landis, J. R., & Koch, G. G. (1977). The Measurement of Observer Agreement for Categorical Data. *Biometrics*, *33*(1), 159–174. http://doi.org/10.2307/2529310

Langfeldt, L. (2001). The Decision-Making Constraints and Processes of Grant Peer Review, and Their Effects on the Review Outcome. *Social Studies of Science*, *31*(6), 820–841. http://doi.org/10.1177/030631201031006002

Langfeldt, L. (2006). The policy challenges of peer review: managing bias, conflict of interests and interdisciplinary assessments. *Research Evaluation*, *15*(1), 31–41. http://doi.org/10.3152/147154406781776039

Latour, B., & Woolgar, S. (1979). *Laboratory life : the social construction of scientific facts*. Beverly Hills: Sage Publications.

Laurent, M. R., & Vickers, T. J. (2009). Seeking Health Information Online: Does Wikipedia Matter? *Journal of the American Medical Informatics Association*, *16*(4), 471–479. http://doi.org/10.1197/jamia.M3059

Leahey, E. (2005). Alphas and Asterisks: The Development of Statistical Significance Testing Standards in Sociology. *Social Forces*, *84*(1), 1–24.

Leahey, E. (2008a). Methodological Memes and Mores: Toward a Sociology of Social Research. *Annual Review of Sociology*, *34*(1), 33–53. http://doi.org/10.1146/annurev.soc.34.040507.134731

Leahey, E. (2008b). Methodological Memes and Mores: Toward a Sociology of Social Research. *Annual Review of Sociology*, *34*(1), 33–53. http://doi.org/10.1146/annurev.soc.34.040507.134731

Leahey, E., & Moody, J. (2014). Sociological Innovation through Subfield Integration. *Social Currents*, *1*(3), 228–256. http://doi.org/10.1177/2329496514540131

Lee, C. J. (2015). Commensuration Bias in Peer Review. *Philosophy of Science*, *82*(5), 1272–1283.

Lee, C. J., Sugimoto, C. R., Zhang, G., & Cronin, B. (2013). Bias in peer review. *Journal of the American Society for Information Science and Technology*, *64*(1), 2–17. http://doi.org/10.1002/asi.22784

Leimu, R., & Koricheva, J. (2005). What determines the citation frequency of ecological papers? *Trends in Ecology & Evolution*, *20*(1), 28–32. http://doi.org/10.1016/j.tree.2004.10.010

Leydesdorff, L. (2009). How are new citation-based journal indicators adding to the bibliometric toolbox? *Journal of the American Society for Information Science and Technology*, *60*(7), 1327–1336. http://doi.org/10.1002/asi.21024

Li, D., & Agha, L. (2015). Big names or big ideas: Do peer-review panels select the best science proposals? *Science*, *348*(6233), 434–438. http://doi.org/10.1126/science.aaa0185

Long, J. B. D., & Lang, K. (1992). Are all Economic Hypotheses False? *Journal of Political Economy*, *100*(6), 1257–1272.

Long, J. S. (1978). Productivity and Academic Position in the Scientific Career. *American Sociological Review*, *43*(6), 889–908. http://doi.org/10.2307/2094628

Lovelock, J. E. (1965). A Physical Basis for Life Detection Experiments. *Nature*, *207*(4997), 568–570. http://doi.org/10.1038/207568a0

Lovelock, J. E. (1990). Hands up for the Gaia hypothesis. *Nature*, *344*(6262), 100–102. http://doi.org/10.1038/344100a0

Luyt, B., & Tan, D. (2010). Improving Wikipedia's credibility: References and citations in a sample of history articles. *Journal of the American Society for Information Science and Technology*, *61*(4), 715–722. http://doi.org/10.1002/asi.21304

Lynch, M., & Bogen, D. (1997). Sociology's Asociological "Core": An Examination of Textbook Sociology in Light of the Sociology of Scientific Knowledge. *American Sociological Review*, *62*(3), 481–493. http://doi.org/10.2307/2657317

Lyons, L. (2013). Discovering the Significance of 5 sigma. *arXiv:1310.1284 [hep-Ex, Physics:hep-Ph, Physics:physics]*. Retrieved from http://arxiv.org/abs/1310.1284

Mahoney, M. J. (1977). Publication prejudices: An experimental study of confirmatory bias in the peer review system. *Cognitive Therapy and Research*, *1*(2), 161–175. http://doi.org/10.1007/BF01173636

Mallard, G., Lamont, M., & Guetzkow, J. (2009). Fairness as Appropriateness Negotiating Epistemological Differences in Peer Review. *Science, Technology & Human Values*, *34*(5), 573–606. http://doi.org/10.1177/0162243908329381

McCloskey, D. N. (1998). *The Rhetoric of Economics* (2 edition). Madison, Wis: University of Wisconsin Press.

McCloskey, D. N., & Ziliak, S. T. (1996). The Standard Error of Regressions. *Journal of Economic Literature*, *34*(1), 97–114.

McPhail, C. (2016, January 7). Personal communication.

Merton, R. K. (1968). The Matthew Effect in Science. *Science*, *159*(3810), 56–&.

Merton, R., & Zuckerman, H. (1971). Patterns of Evaluation in Science: Institutionalisation, Structure and Functions of the Referee System. *Minerva*, *9*(1), 66–100.

Mervis, J. (2014). Peering Into Peer Review. *Science*, *343*(6171), 596–598. http://doi.org/10.1126/science.343.6171.596

Mervis, J. (2015). NIH's peer review stands up to scrutiny. *Science*, *348*(6233), 384–384. http://doi.org/10.1126/science.348.6233.384

Mesgari, M., Okoli, C., Mehdi, M., Nielsen, F. Å., & Lanamäki, A. (2015). "The sum of all human knowledge": A systematic review of scholarly research on the content of Wikipedia. *Journal of the Association for Information Science and Technology*, n/a–n/a. http://doi.org/10.1002/asi.23172

Miller, J. D. (2001). Who is Using the Web for Science and Health Information? *Science Communication*, *22*(3), 256–273. http://doi.org/10.1177/1075547001022003003

Mingers, J., & Leydesdorff, L. (2015). A review of theory and practice in scientometrics. *European Journal of Operational Research*, *246*(1), 1–19. http://doi.org/10.1016/j.ejor.2015.04.002

Moed, H. F. (2007). The effect of "open access" on citation impact: An analysis of ArXiv's condensed matter section. *Journal of the American Society for Information Science and Technology*, *58*(13), 2047–2054. http://doi.org/10.1002/asi.20663

Moore, H. L. (1911). *Laws of wages; an essay in statistical economics.* New York: A.M. Kelley.

Morrison, D. E., & Henkel, R. E. (Eds.). (2006). *The Significance Test Controversy: A Reader* (New edition edition). New Brunswick, N.J: Aldine Transaction.

Myers, G. (1985). Texts as Knowledge Claims: The Social Construction of Two Biology Articles. *Social Studies of Science*, *15*(4), 593–630. http://doi.org/10.1177/030631285015004002

National Academy of Sciences, C. on O. C. and A. (1988). Dictionary of Occupational Titles (DOT). U.S. Dept. of Commerce, Bureau of the Census.

Nelson, J. S. (1990). *Rhetoric Of The Human Sciences: Language And Argument In Scholarship And Public Affairs*. Madison: University of Wisconsin Press.

Newman, M. E. J. (2009). The first-mover advantage in scientific publication. *EPL (Europhysics Letters)*, *86*(6), 68001. http://doi.org/10.1209/0295-5075/86/68001

Nicolaisen, J. (2007). Citation analysis. *Annual Review of Information Science and Technology*, *41*(1), 609–641. http://doi.org/10.1002/aris.2007.1440410120

Nielsen, F. A. (2007). Scientific citations in Wikipedia. *First Monday*, *12*(8). Retrieved from http://firstmonday.org/ojs/index.php/fm/article/view/1997

Nosek, B. A., Spies, J. R., & Motyl, M. (2012). Scientific Utopia II. Restructuring Incentives and Practices to Promote Truth Over Publishability. *Perspectives on Psychological Science*, *7*(6), 615–631. http://doi.org/10.1177/1745691612459058

Nyssa, Z. (2014). *Endangered logics: Conservation science in the American academy* (Ph.D.). The University of Chicago, United States -- Illinois. Retrieved from http://search.proquest.com.proxy.uchicago.edu/pqdtlocal1006268/docview/1620160095/CF277BFB2144BB2PQ/1?accountid=14657

Okoli, C., Mehdi, M., Mesgari, M., Nielsen, F. Å., & Lanamäki, A. (2014). Wikipedia in the eyes of its beholders: A systematic review of scholarly research on Wikipedia readers and readership. *Journal of the Association for Information Science and Technology*, *65*(12), 2381–2403. http://doi.org/10.1002/asi.23162

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716. http://doi.org/10.1126/science.aac4716

Park, H., Lee, J. (Jay), & Kim, B.-C. (2015). Project selection in NIH: A natural experiment from ARRA. *Research Policy*, *44*(6), 1145–1159. http://doi.org/10.1016/j.respol.2015.03.004

Pentland, A. "Sandy." (2012, October 2). Big Data's Biggest Obstacles - HBR. Retrieved December 30, 2014, from https://hbr.org/2012/10/big-datas-biggest-obstacles

Persons, W. M. (1925). Statistics and Economic theory. *The Review of Economics and Statistics*, *7*(3), 179–197. http://doi.org/10.2307/1928417

Pfeffer, J., Leong, A., & Strehl, K. (1977). Paradigm Development and Particularism: Journal Publication in Three Scientific Disciplines. *Social Forces*, *55*(4), 938–951. http://doi.org/10.2307/2577563

Poisson, S. D. (1837). *Recherches sur la probabilité des jugements en matière criminelle et en matière civile: précédées des règles générales du calcul des probabilités*. Bachelier.

Porter, T. M. (1986). *The rise of statistical thinking, 1820-1900*. Princeton, N.J.: Princeton University Press.

Priem, J. (2015). Altmetrics (Chapter from Beyond Bibliometrics: Harnessing Multidimensional Indicators of Scholarly Impact). *arXiv:1507.01328 [cs]*. Retrieved from http://arxiv.org/abs/1507.01328

Quetelet, A. (1835). *Sur l'homme et le développement de ses facultés : ou, Essai de physique sociale*. Paris : Bachelier, imprimeur-libraire, quai des Augustins, no 55. Retrieved from http://archive.org/details/surlhommeetled00quet

Raan, A. F. J. van. (1996). Advanced bibliometric methods as quantitative core of peer review based evaluation and foresight exercises. *Scientometrics*, *36*(3), 397–420. http://doi.org/10.1007/BF02129602

Ransford, H. E., & Miller, J. (1983). Race, Sex and Feminist Outlooks. *American Sociological Review*, *48*(1), 46–59. http://doi.org/10.2307/2095144

Ravitch, D. (1978). The "White Flight" Controversy. *National Affairs*, (51), 135–149.

Rekdal, O. B. (2014). Academic urban legends. *Social Studies of Science*, *44*(4), 638–654. http://doi.org/10.1177/0306312714535679

Rooyen, S. van, Godlee, F., Evans, S., Black, N., & Smith, R. (1999). Effect of open peer review on quality of reviews and on reviewers'recommendations: a randomised trial. *BMJ*, *318*(7175), 23–27. http://doi.org/10.1136/bmj.318.7175.23

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*, *86*(3), 638–641. http://doi.org/10.1037/0033-2909.86.3.638

Roy, R. (1985). Funding Science: The Real Defects of Peer Review and an Alternative to it. *Science, Technology, & Human Values*, *10*(3), 73–81.

Rubin, D. B. (2004). *Multiple Imputation for Nonresponse in Surveys*. John Wiley & Sons.

Rueschemeyer, D. (2009). *Usable Theory: Analytic Tools for Social and Political Research* (First Edition edition). Princeton, N.J: Princeton University Press.

Rzhetsky, A., Shatkay, H., & Wilbur, W. J. (2009). How to get the most out of your curation effort. *PLoS Comput Biol*, *5*(5), e1000391. http://doi.org/10.1371/journal.pcbi.1000391

Sabaj Meruane, O., González Vergara, C., & Pina-Stranger, Á. (2016). What We Still Don't Know About Peer Review. *Journal of Scholarly Publishing*, *47*(2), 180–212. http://doi.org/10.3138/jsp.47.2.180

Samoilenko, A., & Yasseri, T. (2014). The distorted mirror of Wikipedia: a quantitative analysis of Wikipedia coverage of academics. *EPJ Data Science*, *3*(1). http://doi.org/10.1140/epjds20

Sample, I. (2012, April 12). Harvard University says it can't afford journal publishers' prices. Retrieved June 16, 2015, from http://www.theguardian.com/science/2012/apr/24/harvard-university-journal-publishers-prices

Schroeder, R., & Taylor, L. (2015). Big data and Wikipedia research: social science knowledge across disciplinary divides. *Information, Communication & Society*, *0*(0), 1–18. http://doi.org/10.1080/1369118X.2015.1008538

Seglen, P. O. (1997). Why the impact factor of journals should not be used for evaluating research. *BMJ*, *314*(7079), 497. http://doi.org/10.1136/bmj.314.7079.497

Shapin, S. (1984). Pump and Circumstance: Robert Boyle's Literary Technology. *Social Studies of Science*, *14*(4), 481–520.

Shaw, A., & Hill, B. M. (2014). Laboratories of Oligarchy? How the Iron Law Extends to Peer Production. *Journal of Communication*, *64*(2), 215–238. http://doi.org/10.1111/jcom.12082

Shuai, X., Jiang, Z., Liu, X., & Bollen, J. (2013). A Comparative Study of Academic and Wikipedia Ranking. In *Proceedings of the 13th ACM/IEEE-CS Joint Conference on Digital Libraries* (pp. 25–28). New York, NY, USA: ACM. http://doi.org/10.1145/2467696.2467746

Siler, K., Lee, K., & Bero, L. (2015). Measuring the effectiveness of scientific gatekeeping. *Proceedings of the National Academy of Sciences*, *112*(2), 360–365. http://doi.org/10.1073/pnas.1418218112

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-Positive Psychology Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant. *Psychological Science*, 0956797611417632. http://doi.org/10.1177/0956797611417632

Simon, R. J., Bakanic, V., & McPhail, C. (1986). Who Complains to Journal Editors and What Happens*. *Sociological Inquiry*, *56*(2), 259–271. http://doi.org/10.1111/j.1475-682X.1986.tb00087.x

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). P-curve: A key to the file-drawer. *Journal of Experimental Psychology: General*, *143*(2), 534–547. http://doi.org/10.1037/a0033242

Smelser, N. J. (2015). Sources of Unity and Disunity in Sociology. *The American Sociologist*, *46*(3), 303–312. http://doi.org/10.1007/s12108-015-9260-2

Sokal, M. M. (1987). James McKeen Cattell and Mental Anthropometry. In M. M. Sokal (Ed.), *Psychological testing and American society, 1890-1930*. New Brunswick: Rutgers University Press.

Spoerri, A. (2007). What is popular on Wikipedia and why? *First Monday*, *12*(4). http://doi.org/10.5210/fm.v12i4.1765

Steiner, T. (2014). Bots vs. Wikipedians, Anons vs. Logged-Ins (Redux): A Global Study of Edit Activity on Wikipedia and Wikidata. In *Proceedings of The International Symposium on Open Collaboration* (pp. 25:1–25:7). New York, NY, USA: ACM. http://doi.org/10.1145/2641580.2641613

Sterling, T. D. (1959). Publication decisions and their possible effects on inferences drawn from tests of significance—or vice versa. *Journal of the American Statistical Association*, *54*, 30–34. http://doi.org/10.2307/2282137

Stigler, G. J. (1962). Henry L. Moore and Statistical Economics. *Econometrica*, *30*(1), 1–21. http://doi.org/10.2307/1911284

Stigler, S. M. (1978). Francis Ysidro Edgeworth, Statistician. *Journal of the Royal Statistical Society. Series A (General)*, *141*(3), 287–322. http://doi.org/10.2307/2344804

Stigler, S. M. (1986). *The history of statistics : the measurement of uncertainty before 1900*. Cambridge, Mass.: Belknap Press of Harvard University Press.

Stigler, S. M. (1989). Francis Galton's Account of the Invention of Correlation. *Statistical Science*, *4*(2), 73–79.

Stocking, G. W. (1968). Franz Boaz and the Culture Concept in Historical Perspective. In *Race, culture, and evolution; essays in the history of anthropology* (pp. 161–194). New York: Free Press.

Strang, D., & Siler, K. (2015). Revising as Reframing Original Submissions versus Published Papers in Administrative Science Quarterly, 2005 to 2009. *Sociological Theory*, *33*(1), 71–96. http://doi.org/10.1177/0735275115572152

Tam, T. (1997). Sex Segregation and Occupational Gender Inequality in the United States: Devaluation or Specialized Training? *American Journal of Sociology*, *102*(6), 1652–1692. http://doi.org/10.1086/231129

Tam, T. (2000). Occupational Wage Inequality and Devaluation: A Cautionary Tale of Measurement Error. *American Journal of Sociology*, *105*(6), 1752–1760. http://doi.org/10.1086/210472

Teplitskiy, M. (2015a). Frame Search and Re-search: How Quantitative Sociological Articles Change During Peer Review. *The American Sociologist*, 1–25. http://doi.org/10.1007/s12108-015-9288-3

Teplitskiy, M. (2015b). Frame Search and Re-Search: How Quantitative Sociological Articles Change During Peer Review. *Available at SSRN*. Retrieved from http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2634766

Teplitskiy, M., & Bakanic, V. (2016). Do Peer Reviews Predict Impact? Evidence from the American Sociological Review, 1978 to 1982. *Socius: Sociological Research for a Dynamic World*, *2*, 2378023116640278. http://doi.org/10.1177/2378023116640278

The CMS Collaboration. (2012). Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC. *Physics Letters B*, *716*(1), 30–61. http://doi.org/10.1016/j.physletb.2012.08.021

The Economist. (2013, October 19). Trouble at the lab. *The Economist*. Retrieved from http://www.economist.com/news/briefing/21588057-scientists-think-science-self-correcting-alarming-degree-it-not-trouble

The Open Access Citation Advantage Service. (n.d.). Retrieved October 9, 2015, from http://sparceurope.org/oaca/

Travis, G. D. L., & Collins, H. M. (1991). New Light on Old Boys: Cognitive and Institutional Particularism in the Peer Review System. *Science, Technology & Human Values*, *16*(3), 322–341. http://doi.org/10.1177/016224399101600303

Trench, B. (2008). Internet: turning science communication inside-out? In M. Bucchi & B. Trench (Eds.), *Handbook of Public Communication of Science and Technology*. London and New York: Routledge. Retrieved from http://www.routledgesociology.com/books/Handbook-of-Public-Communication-of-Science-and-Technology-isbn9780415386173

Udell, M., Horn, C., Zadeh, R., & Boyd, S. (2014). Generalized Low Rank Models. *arXiv:1410.0342 [cs, Math, Stat]*. Retrieved from http://arxiv.org/abs/1410.0342

Van den Besselaar, P., & Sandström, U. (2015). Early career grants, performance, and careers: A study on predictive validity of grant decisions. *Journal of Informetrics*, *9*(4), 826–838. http://doi.org/10.1016/j.joi.2015.07.011

Van Noorden, R. (2013). Open access: The true cost of science publishing. *Nature*, *495*(7442), 426–429. http://doi.org/10.1038/495426a

Van Noorden, R., Maher, B., & Nuzzo, R. (2014). The top 100 papers. *Nature*, *514*(7524), 550–553. http://doi.org/10.1038/514550a

Waltman, L. (2015). A review of the literature on citation impact indicators. *arXiv:1507.02099 [cs]*. Retrieved from http://arxiv.org/abs/1507.02099

West, R., Weber, I., & Castillo, C. (2012). Drawing a Data-driven Portrait of Wikipedia Editors. In *Proceedings of the Eighth Annual International Symposium on Wikis and Open Collaboration* (pp. 3:1–3:10). New York, NY, USA: ACM. http://doi.org/10.1145/2462932.2462937

Wetzels, R., Matzke, D., Lee, M. D., Rouder, J. N., Iverson, G. J., & Wagenmakers, E.-J. (2011). Statistical Evidence in Experimental Psychology An Empirical Comparison Using 855 t Tests. *Perspectives on Psychological Science*, *6*(3), 291–298. http://doi.org/10.1177/1745691611406923

Wren, J. D. (2005a). Open access and openly accessible: a study of scientific publications shared via the internet. *BMJ*, *330*(7500), 1128. http://doi.org/10.1136/bmj.38422.611736.E0

Wren, J. D. (2005b). Open access and openly accessible: a study of scientific publications shared via the internet. *BMJ*, *330*(7500), 1128. http://doi.org/10.1136/bmj.38422.611736.E0

Wuchty, S., Jones, B. F., & Uzzi, B. (2007). The increasing dominance of teams in production of knowledge. *Science*, *316*(5827), 1036–9. http://doi.org/10.1126/science.1136099

Wundt, W. (1862). Contributions to the Theory of Sensory Perception. In *Classics in psychology* (Vol. xx, pp. 51–78). Oxford, England: Philosophical Library.

Xie, Y. (1988). Franz Boas and Statistics. *Annals of Scholarship*, *5*, 269–96.

Young, C. (2009). Model Uncertainty in Sociological Research: An Application to Religion and Economic Growth. *American Sociological Review*, *74*(3), 380–397.

Zuckerman, H., & Merton, R. K. (1971). Patterns of evaluation in science: Institutionalisation, structure and functions of the referee system. *Minerva*, *9*(1), 66–100. http://doi.org/10.1007/BF01553188