How Firm is Sociological Knowledge?

Reanalysis of GSS findings with alternative models and out-of-sample data, 1972-2012

James Evans and Misha Teplitskiy
Department of Sociology
University of Chicago

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, and Wittenberg 2000; McCloskey and Ziliak 1996; Morrison and 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 “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

[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.

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 nonrobust 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 (Ioannidis et al. 2001), machine learning (Pentland 2012) and high energy physics (Lyons 2013). A featured cover article from *The Economist* titled "Unreliable Research: Trouble at the lab" (The Economist 2013) and several related popular 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 (Lovelock 1965, 1990), 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 and 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; 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[4] (Beirne 1987; Quetelet 1835). This average-based or "Continental approach" to statistics (Duncan 1984) entered German experimental psychology (Gigerenzer 1987; Porter 1986; Stigler 1986; Wundt 1862:71–72) 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 and Xie 1994; Camic 1995).[5]

The "science of variation" or "British approach" to statistical inference (Camic and 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 1986:129). He was not interested in the average man, but "men…different from the average"—those of superior and inferior

[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 Quetelet's measures were "simple statistics", which tarnished his own earlier use of "social physics."

[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).

[4] August Compte coined "sociologie" to reject this approach to measurement. He quipped that Quetelet's measures were "simple statistics" that tarnished his own earlier use of "social physics" (Beirne 1987; Comte 1830).

[5] 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" (1899:151–52). This category-centered view made categorical membership data central, and so appropriate sociological data were typically discrete or "absolute numbers" (1901:29, 1910:722; Camic and Xie 1994).

human endowments and the hereditary basis of their differences (Hilts 1973:229). 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; 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 (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 on averages (Boas 1894:227, 1897:151; Sokal 1987; Stocking 1968).[6] 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 and Xie 1994; Moore 1911:19). Moore was first to "furnish empirical estimates of the parameters in theoretical models" by statistically controlling for confounding factors in natural settings (Moore 1911:23; Stigler 1962; Camic and 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

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

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 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.[7]

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 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

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

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 and Rosenfeld 1968), or from a *p*-value threshold of .05 to .003.[8] 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

[8] Personal communication with Tom Witten, 2010.

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).[9] 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 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.

[9] 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).

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 and Tamhane 1992), or related approaches that explicitly control for the expected proportion of falsely rejected hypotheses or false discovery rate (Benjamini and Hochberg 1995; Barber and 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 and 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, and 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.[10] Such robustness exchanges are rare in the sociological literature.

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*.

[10] 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).

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.[11] These efforts produced a database with metadata describing thousands of social science articles, books, and other publications. Andrew Abbott used these data in his widely cited article, "Seven types of ambiguity" in order to explore the diverse readings of an indicator capturing strength of religious attachment (2013).

We supplemented these metadata with information describing in more detail the data analysis performed in each publication. For the subset of articles containing

[11] 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).

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, and Wilbur 2009). See appendix and Table A2 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.[12] 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°.

$$\mathrm{Y}_i = \beta_0 + \sum \beta_j \mathrm{X}_i^* + \sum \beta_k \mathrm{X}_i^{\circ} + \varepsilon$$

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[13]. Details of how models were estimated and some illustrative examples of published papers, for which this estimation is a reasonable approximation, are found in the Appendix (see Figure A4).

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. 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

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

[13] 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).

enables us to test the robustness of findings at time of publication and several decades afterward.

Figure 1 about here

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[14]. 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 1 presents the most common *original variable* → *cognate variable* substitutions.

Table 1 about here

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

In addition, we evaluate robustness at time of publication by estimating models on waves of the GSS that came immediately after the latest wave the article used. 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.

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[15] 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*[16]—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.

[15] 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.

[16] 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.

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).

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 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 2 about here

The top panel of Figure 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 3 about here

Figure 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 1).

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[17]. The sample for this analysis draws on 398 qualifying articles. Figure 4 summarizes the consequences of this perturbation.

[17] 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 about here

Figure 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 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 5 about here

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 more sparse. 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 A3.).

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 (Ioannidis 2005; Long and Lang 1992; Simmons et al. 2011; Simonsohn, Nelson, and 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 A1 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 2008), evaluation (Lamont 2009, 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.

References

Abbott, Andrew. 1988. "Transcending General Linear Reality." *Sociological Theory* 6(2):169–86.

Abbott, Andrew. 2013. "Seven Types of Ambiguity." *Theory and Society* 26(2-3):357–99.

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). Retrieved May 7, 2015 (http://arxiv.org/abs/1207.0319).

Babbie, Earl R. 2014. *The Practice of Social Research*.

Baltay, Charles and Arthur H. Rosenfeld. 1968. *Meson Spectroscopy; a Collection of Articles*. New York: W. A. Benjamin. Retrieved January 7, 2015 (http://pi.lib.uchicago.edu/1001/cat/bib/1082525).

Barber, Rina Foygel and Emmanuel Candes. 2014. "Controlling the False Discovery Rate via Knockoffs." *arXiv:1404.5609 [math, stat]*. Retrieved January 15, 2015 (http://arxiv.org/abs/1404.5609).

Beirne, Piers. 1987. "Adolphe Quetelet and the Origins of Positivist Criminology." *American Journal of Sociology* 92(5):1140–69.

Benjamini, Yoav and Yosef Hochberg. 1995. "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing." *Journal of the Royal Statistical Society. Series B (Methodological)* 57(1):289–300.

Boas, Franz. 1894. "Human Faculty as Determined by Race." Pp. 221–42 in *The shaping of American anthropology, 1883-1911; a Franz Boas reader.*, edited by G. W. Stocking. New York: Basic Books.

Boas, Franz. 1897. *Race, Language and Culture*. New York: The Macmillan Co.

Bonferroni, Carlo Emilio. 1936. *Teoria statistica delle classi e calcolo delle probabilità*. Firenze: Seeber.

Button, Katherine S. et al. 2013. "Power Failure: Why Small Sample Size Undermines the Reliability of Neuroscience." *Nature Reviews Neuroscience* 14(5):365–76.

Camic, Charles. 1995. "Three Departments in Search of a Discipline: Localism and Interdisciplinary Interaction in American Sociology, 1890—1940." *Social Research* 62(4):1003–33.

Camic, Charles and Yu Xie. 1994. "The Statistical Turn in American Social Science: Columbia University, 1890 to 1915." *American Sociological Review* 59(5):773–805.

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

Cole, Stephen, ed. 2001. *What's Wrong with Sociology?* New Brunswick, N.J: Transaction Publishers.

Coleman, James S. et al. 1966. *Equality of Educational Opportunity Study*. U.S. Department of Health, Education and Welfare.

Comte, Auguste. 1830. *Cours de philosophie positive. [Tome 4] / par M. Auguste Comte,...* Rouen frères (Bachelier) (Paris). Retrieved December 31, 2014 (http://gallica.bnf.fr/ark:/12148/bpt6k76270k).

Dudoit, Sandrine and Mark J. Van der Laan. 2008. *Multiple Testing Procedures with Applications to Genomics*. Springer. Retrieved January 15, 2015 (http://www.springer.com/life+sciences/biochemistry+%26+biophysics/book/978-0-387-49316-9).

Duncan, Otis Dudley. 1984. *Notes on Social Measurement: Historical and Critical*. Russell Sage Foundation.

Dunn, Olive Jean. 1959. "Estimation of the Medians for Dependent Variables." *The Annals of Mathematical Statistics* 30(1):192–97.

Dunn, Olive Jean. 1961. "Multiple Comparisons among Means." *Journal of the American Statistical Association* 56(293):52–64.

Dunnett, Charles W. and Ajit C. Tamhane. 1992. "A Step-Up Multiple Test Procedure." *Journal of the American Statistical Association* 87(417):162–70.

England, Paula, Joan M. Hermsen, and David A. Cotter. 2000. "The Devaluation of Women's Work: A Comment on Tam." *American Journal of Sociology* 105(6):1741–51.

Fleiss, J. L. 1981. *Statistical Methods for Rates and Proportions*. Second edition. New York: John Wiley.

Fox, Mary F. and Glenn Firebaugh. 1992. "Confidence in Science: The Gender Gap." *Social Science Quarterly* 73(1):101–13.

Franco, Annie, Neil Malhotra, and Gabor Simonovits. 2014. "Publication Bias in the Social Sciences: Unlocking the File Drawer." *Science* 345(6203):1502–5.

Freese, Jeremy. 2007. "Replication Standards for Quantitative Social Science Why Not Sociology?" *Sociological Methods & Research* 36(2):153–72.

Gerber, A. and N. Malhotra. 2008. “Do Statistical Reporting Standards Affect What Is Published? Publication Bias in Two Leading Political Science Journals.” *Quarterly Journal of Political Science* 3(3):313–26.

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

Gibson, Lydialyle. 2013. “Growing Numbers.” *University of Chicago Magazine*. Retrieved March 3, 2014 (http://mag.uchicago.edu/law-policy-society/growing-numbers?msource=MAG10).

Giddens, Franklin. 1899. “Exact Methods in Sociology.” *Popular Science Monthly* 56:145–59.

Giddings, Franklin H. 1910. “The Social Marking System.” *American Journal of Sociology* 15(6):721–40.

Giddings, Franklin Henry. 1901. *Inductive Sociology; a Syllabus of Methods, Analyses and Classifications, and Provisionally Formulated Laws.* New York; London: The Macmillan company; Macmillan & Co., Ltd.

Gigerenzer, Gerd. 1987. “The Probabilistic Revolution in Psychology--An Overview.” Pp. 7–9 in *The probabilistic revolution, Vol. 1: Ideas in history; Vol. 2: Ideas in the sciences*. Cambridge, MA, US: The MIT Press.

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

Hilts, Victor L. 1973. “Statistics and Social Science.” in *Foundations of scientific method: the nineteenth century.*, edited by R. N. Giere and R. S. Westfall. Bloomington: Indiana University Press.

Hilts, Victor L. 1981. *Statist and Statistician*. New York: Arno Press.

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

Hubbard, Raymond. 2004. “Alphabet Soup Blurring the Distinctions Betweenp’s Anda’s in Psychological Research.” *Theory & Psychology* 14(3):295–327.

Ioannidis, J. P. 2005. “Why Most Published Research Findings Are False.” *PLoS Med* 2(8):e124.

Ioannidis, John and Chris Doucouliagos. 2013. “What’s to Know About the Credibility of Empirical Economics?” *Journal of Economic Surveys* 27(5):997–1004.

Ioannidis, John P. A., Evangelia E. Ntzani, Thomas A. Trikalinos, and Despina G. Contopoulos-Ioannidis. 2001. "Replication Validity of Genetic Association Studies." *Nature Genetics* 29(3):306–9.

Jeřábek, Hynek. 2001. "Paul Lazarsfeld—The Founder of Modern Empirical Sociology: A Research Biography." *International Journal of Public Opinion Research* 13(3):229–44.

King, Gary, Michael Tomz, and Jason Wittenberg. 2000. "Making the Most of Statistical Analyses: Improving Interpretation and Presentation." *American Journal of Political Science* 44(2):347–61.

Knorr-Cetina, K. 1999. *Epistemic Cultures: How the Sciences Make Knowledge*. Cambridge, Mass.: Harvard University Press.

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

Lamont, Michèle. 2009. *How Professors Think : Inside the Curious World of Academic Judgment*. Cambridge, Mass.: Harvard University Press.

Lamont, Michèle. 2012. "Toward a Comparative Sociology of Valuation and Evaluation." *Annual Review of Sociology* 38(1):201–21.

Landis, J. Richard and Gary G. Koch. 1977. "The Measurement of Observer Agreement for Categorical Data." *Biometrics* 33(1):159–74.

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

Leahey, Erin. 2008. "Methodological Memes and Mores: Toward a Sociology of Social Research." *Annual Review of Sociology* 34(1):33–53.

Long, J. Bradford De and Kevin Lang. 1992. "Are All Economic Hypotheses False?" *Journal of Political Economy* 100(6):1257–72.

Lovelock, J. E. 1965. "A Physical Basis for Life Detection Experiments." *Nature* 207(4997):568–70.

Lovelock, James E. 1990. "Hands up for the Gaia Hypothesis." *Nature* 344(6262):100–102.

Lyons, Louis. 2013. "Discovering the Significance of 5 Sigma." *arXiv:1310.1284 [hep-ex, physics:hep-ph, physics:physics]*. Retrieved December 30, 2014 (http://arxiv.org/abs/1310.1284).

McCloskey, Deirdre N. and Stephen T. Ziliak. 1996. “The Standard Error of Regressions.” *Journal of Economic Literature* 34(1):97–114.

Moore, Henry Ludwell. 1911. *Laws of Wages; an Essay in Statistical Economics.* New York: A.M. Kelley.

Morrison, Denton E. and Ramon E. Henkel, eds. 2006. *The Significance Test Controversy: A Reader*. New edition edition. New Brunswick, N.J: Aldine Transaction.

National Academy of Sciences, Committee on Occupational Classification and Analysis. 1988. “Dictionary of Occupational Titles (DOT).”

Nyssa, Zoe. 2014. “Endangered Logics: Conservation Science in the American Academy.” Ph.D., The University of Chicago, United States -- Illinois. Retrieved December 31, 2014 (http://search.proquest.com.proxy.uchicago.edu/pqdtlocal1006268/docview/1620160095/CF277BFB2144BB2PQ/1?accountid=14657).

Pentland, Alex “Sandy.”2012. “Big Data’s Biggest Obstacles - HBR.” *Harvard Business Review*. Retrieved December 30, 2014 (https://hbr.org/2012/10/big-datas-biggest-obstacles).

Persons, Warren M. 1925. “Statistics and Economic Theory.” *The Review of Economics and Statistics* 7(3):179–97.

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

Porter, Theodore M. 1986. *The Rise of Statistical Thinking, 1820-1900*. Princeton, N.J.: Princeton University Press.

Quetelet, Adolphe. 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 December 31, 2014 (http://archive.org/details/surlhommeetled00quet).

Ransford, H. Edward and Jon Miller. 1983. “Race, Sex and Feminist Outlooks.” *American Sociological Review* 48(1):46–59.

Ravitch, Diane. 1978. “The ‘White Flight’ Controversy.” *National Affairs* (51):135–49.

Rosenthal, Robert. 1979. “The File Drawer Problem and Tolerance for Null Results.” *Psychological Bulletin* 86(3):638–41.

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

Rzhetsky, A., H. Shatkay, and W. J. Wilbur. 2009. "How to Get the Most out of Your Curation Effort." *PLoS Comput Biol* 5(5):e1000391.

Simmons, Joseph P., Leif D. Nelson, and Uri Simonsohn. 2011. "False-Positive Psychology Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant." *Psychological Science* 0956797611417632.

Simonsohn, Uri, Leif D. Nelson, and Joseph P. Simmons. 2014. "P-Curve: A Key to the File-Drawer." *Journal of Experimental Psychology: General* 143(2):534–47.

Sokal, Michael M. 1987. "James McKeen Cattell and Mental Anthropometry." in *Psychological testing and American society, 1890-1930*, edited by M. M. Sokal. New Brunswick: Rutgers University Press.

Stigler, George J. 1962. "Henry L. Moore and Statistical Economics." *Econometrica* 30(1):1–21.

Stigler, Stephen M. 1978. "Francis Ysidro Edgeworth, Statistician." *Journal of the Royal Statistical Society. Series A (General)* 141(3):287–322.

Stigler, Stephen M. 1986. *The History of Statistics : The Measurement of Uncertainty before 1900*. Cambridge, Mass.: Belknap Press of Harvard University Press.

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

Stocking, George W. 1968. "Franz Boaz and the Culture Concept in Historical Perspective." Pp. 161–94 in *Race, culture, and evolution; essays in the history of anthropology*. New York: Free Press.

Tam, Tony. 1997. "Sex Segregation and Occupational Gender Inequality in the United States: Devaluation or Specialized Training?" *American Journal of Sociology* 102(6):1652–92.

Tam, Tony. 2000. "Occupational Wage Inequality and Devaluation: A Cautionary Tale of Measurement Error." *American Journal of Sociology* 105(6):1752–60.

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.

The Economist. 2013. "Trouble at the Lab." *The Economist*, October 19. Retrieved December 30, 2014 (http://www.economist.com/news/briefing/21588057-scientists-think-science-self-correcting-alarming-degree-it-not-trouble).

Udell, Madeleine, Corinne Horn, Reza Zadeh, and Stephen Boyd. 2014. "Generalized Low Rank Models." *arXiv:1410.0342 [cs, math, stat]*. Retrieved May 7, 2015 (http://arxiv.org/abs/1410.0342).

Wetzels, Ruud et al. 2011. “Statistical Evidence in Experimental Psychology An Empirical Comparison Using 855 T Tests.” *Perspectives on Psychological Science* 6(3):291–98.

Wundt, Wilhelm. 1862. “Contributions to the Theory of Sensory Perception.” Pp. 51–78 in *Classics in psychology*, vol. xx. Oxford, England: Philosophical Library.

Xie, Yu. 1988. “Franz Boas and Statistics.” *Annals of Scholarship* 5:269–96.

Young, Cristobal. 2009. “Model Uncertainty in Sociological Research: An Application to Religion and Economic Growth.” *American Sociological Review* 74(3):380–97.

Appendix

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 (Fox and 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 A1.

Table A1 about here

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.[18] Coding was performed through a website that allowed students access to the digital article. Table A1 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 (1981:218). 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 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

[18] 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.

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 A2 about here

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 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. One of the authors co-authored the open source pyanno software that implements this model and is available for download at: https://pypi.python.org/pypi/pyanno/2.0.2.

Data Description

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

Figure A1 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 A1 about here

Figure A2 below displays time-trends in the average number of dependent and independent variables per article published in a given year.

Figure A2 about here

Figure A3 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 A3 about here

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 below illustrates this approach and provides examples of two publications from the data sample.

Figure A4 about here

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 et al. 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 present models with several dependent variables, several independent variables, and estimate these models on several years of data (see Figure A2). 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.

Table 1a: 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

Table 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 1/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)

GSS Regression Measures

$$\underbrace{y_i - \epsilon_i}_{\text{Model Fit}} = \beta_0 + \underbrace{\underbrace{\beta_1 x^*_{i,1} + ... + \beta_m x^*_{i,m}}_{\text{Central Variables}} + \beta_{m+1} x^\circ_{m+1} + ... + \beta_n x^\circ_{i,n}}_{\text{All Variables}}$$

Model Fit

$$R^2 = \frac{\sum_{i=1}^N \epsilon_i^2}{\sum_{i=1}^N \left(y_i - \frac{1}{N}\sum_{i=1}^N y_i\right)^2}$$

$$\bar{R}^2 = 1 - \left(1 - R^2\right)\frac{N-1}{N-n-1}$$

Central Variables

$$\sum_{j=1}^{m} \delta_j \begin{cases} 0 & \text{if } \beta_j, p > .05 \\ 1 & \text{if } \beta^*_j, p < .05 \end{cases} = \textit{significant effects} = \sum_{j=1}^{n} \delta_j \begin{cases} 0 & \text{if } \beta_j, p > .05 \\ 1 & \text{if } \beta^*_j, p < .05 \end{cases}$$

All Variables

$$\frac{\sum_{j=1}^m |\beta_j|}{m} = \textit{effect size} = \frac{\sum_{j=1}^n |\beta_j|}{n}$$

Figure 1. Measurement strategy for capturing model fit and significance across original and perturbed models

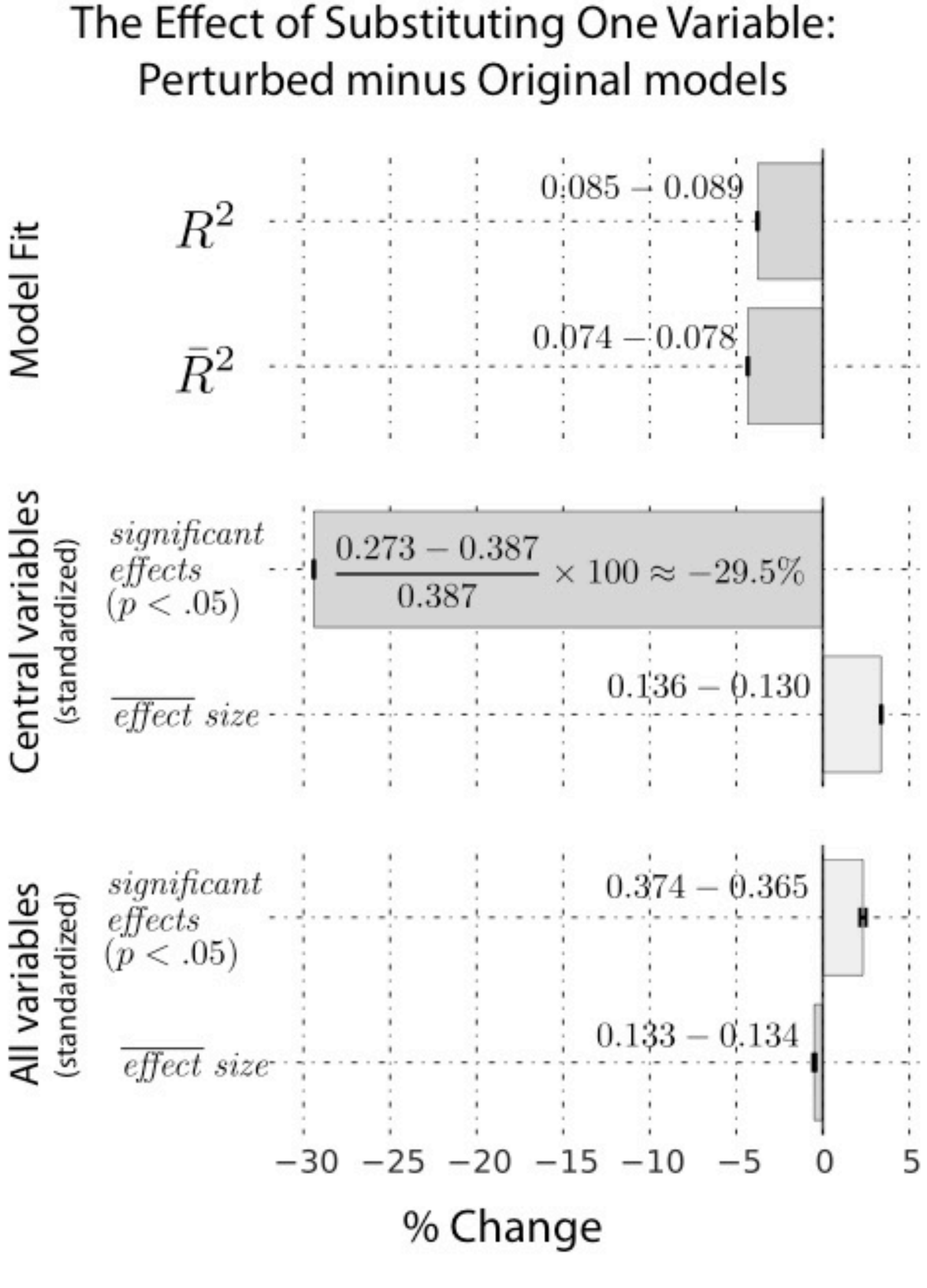

Figure 2. Change in model fit after the original model was perturbed by the substitution of one randomly selected central variable. One of the bars shows the calculation of percent change of mean outcome from the original to the perturbed models. The numbers next to the other bars show the mean outcomes of the perturbed and original models.

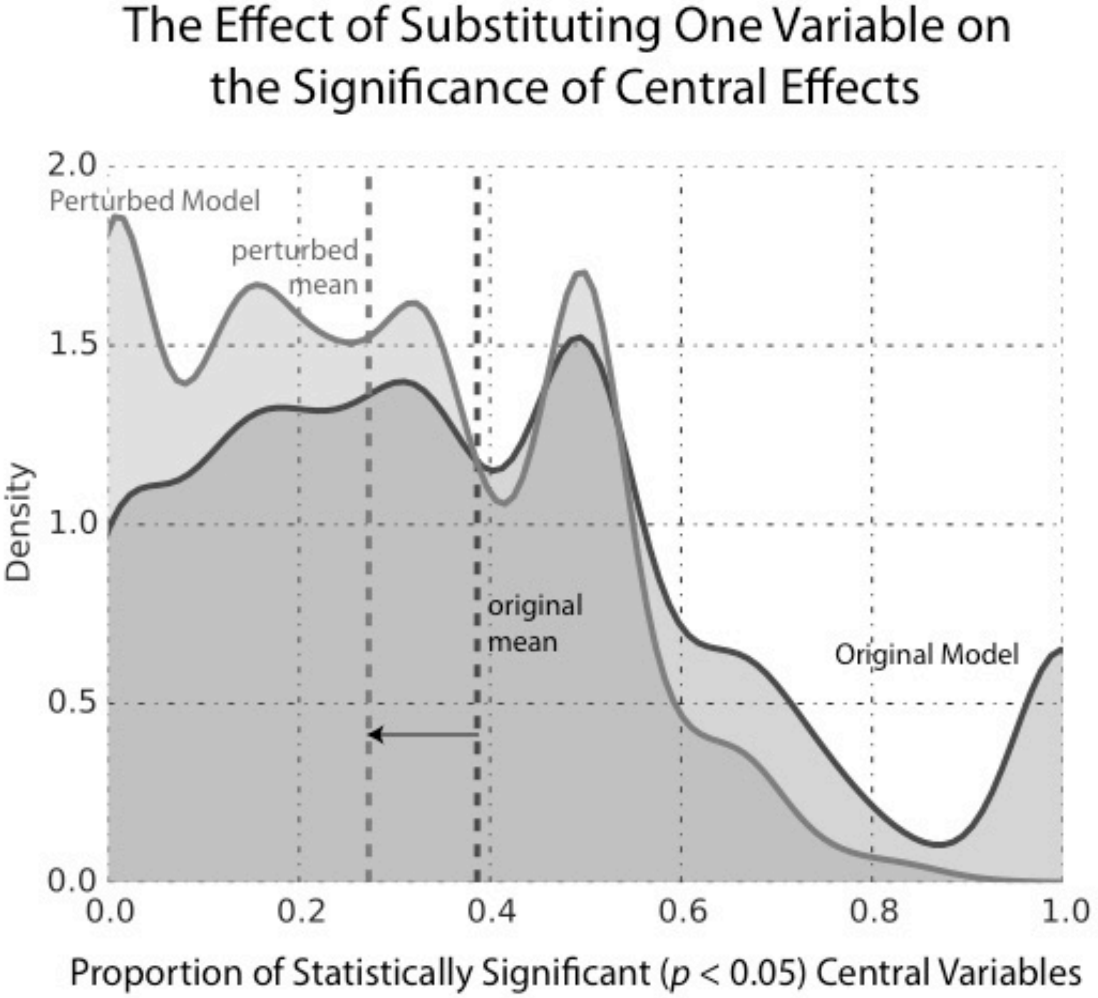

Figure 3. Shift in the distribution of significant central variables with the substitution of one, randomly selected variable.

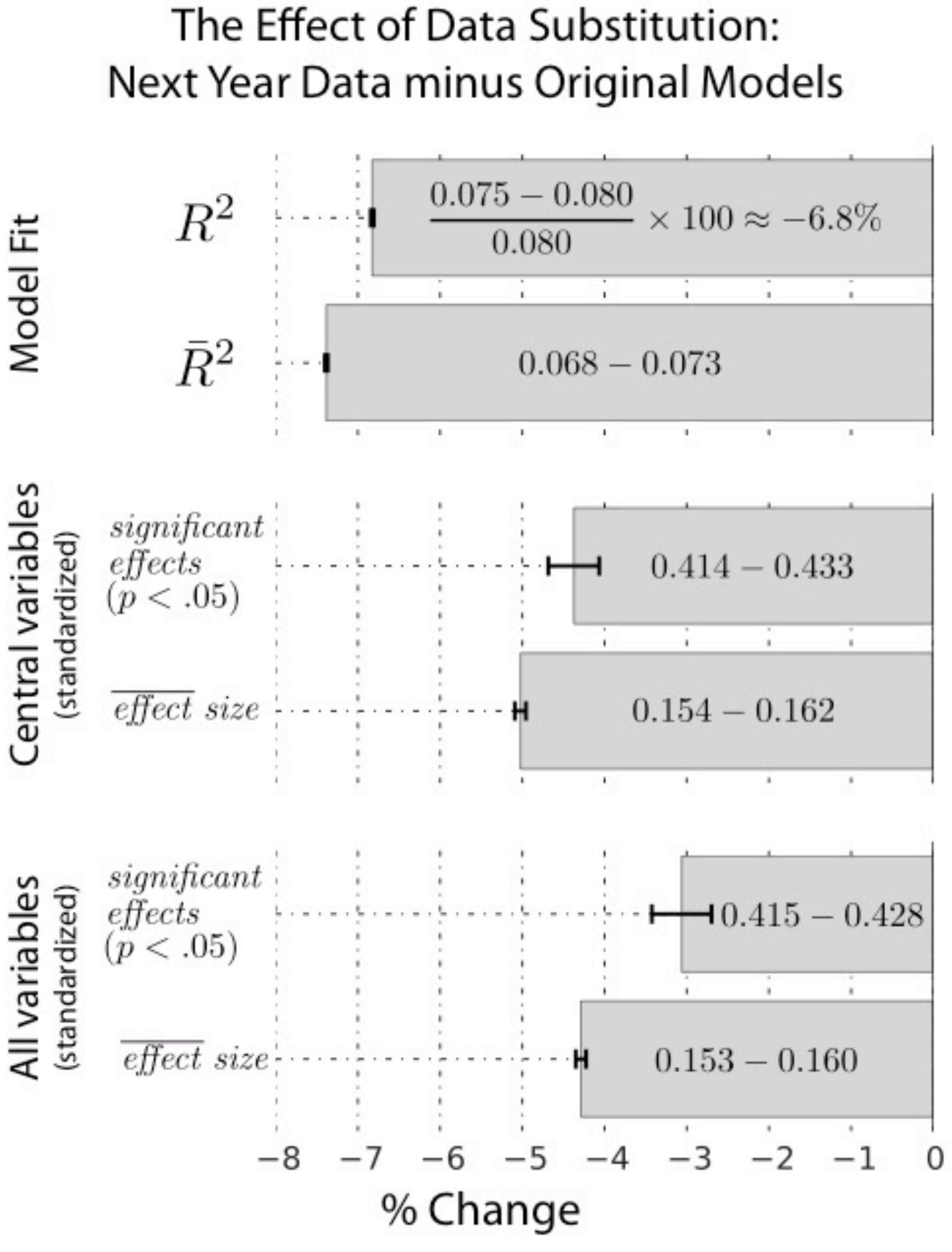

Figure 4. Change in model fit after the original model was re-estimated on data the next available year after publication. One of the bars shows the calculation of percent change of mean outcome from the original to the perturbed models. The numbers next to the other bars show the mean outcomes of the perturbed and original models.

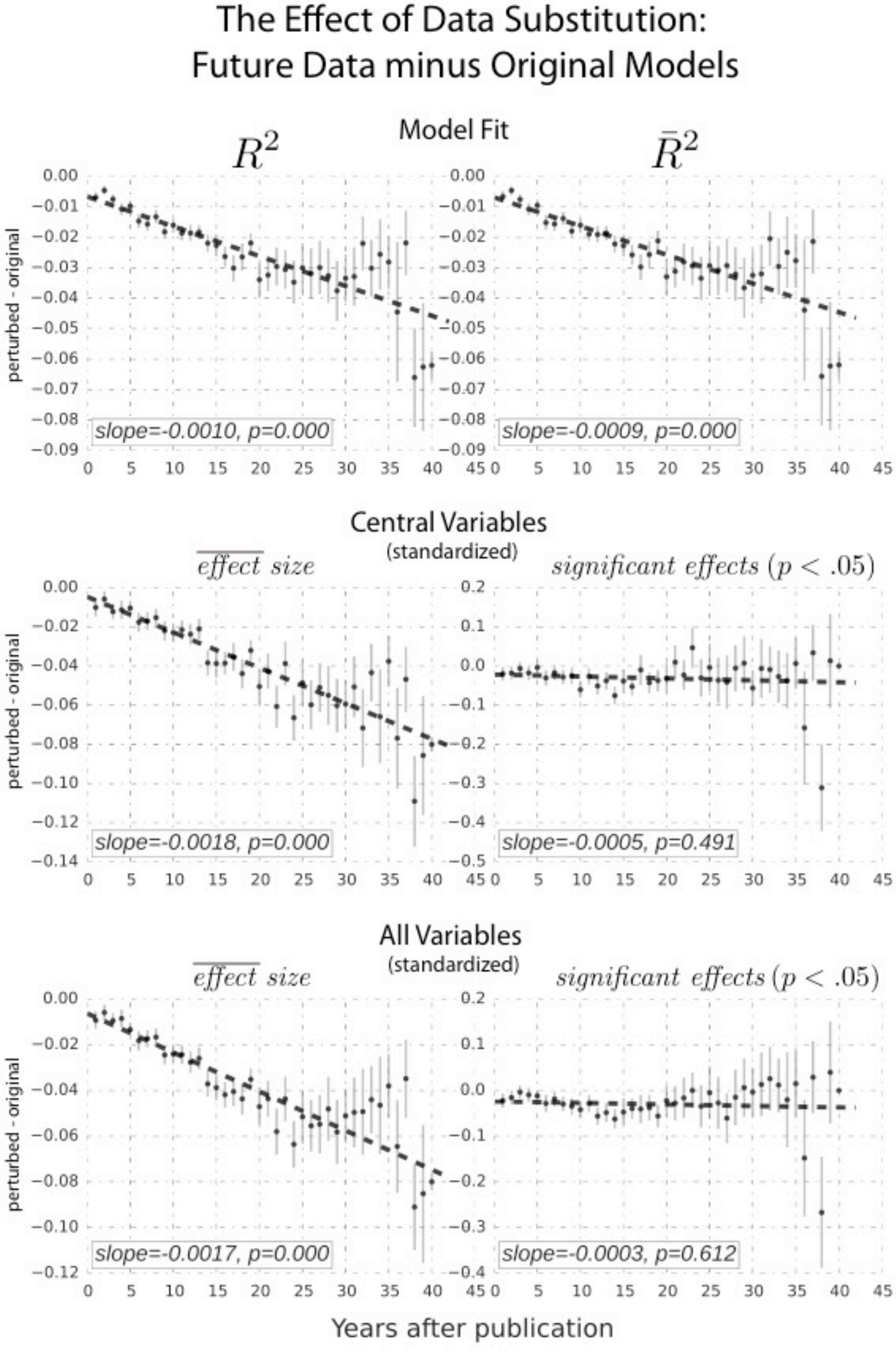

Figure 5. Change in model fit after the original model was re-estimated on data in each available year following publication.

Table A1. 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

Table A2. 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

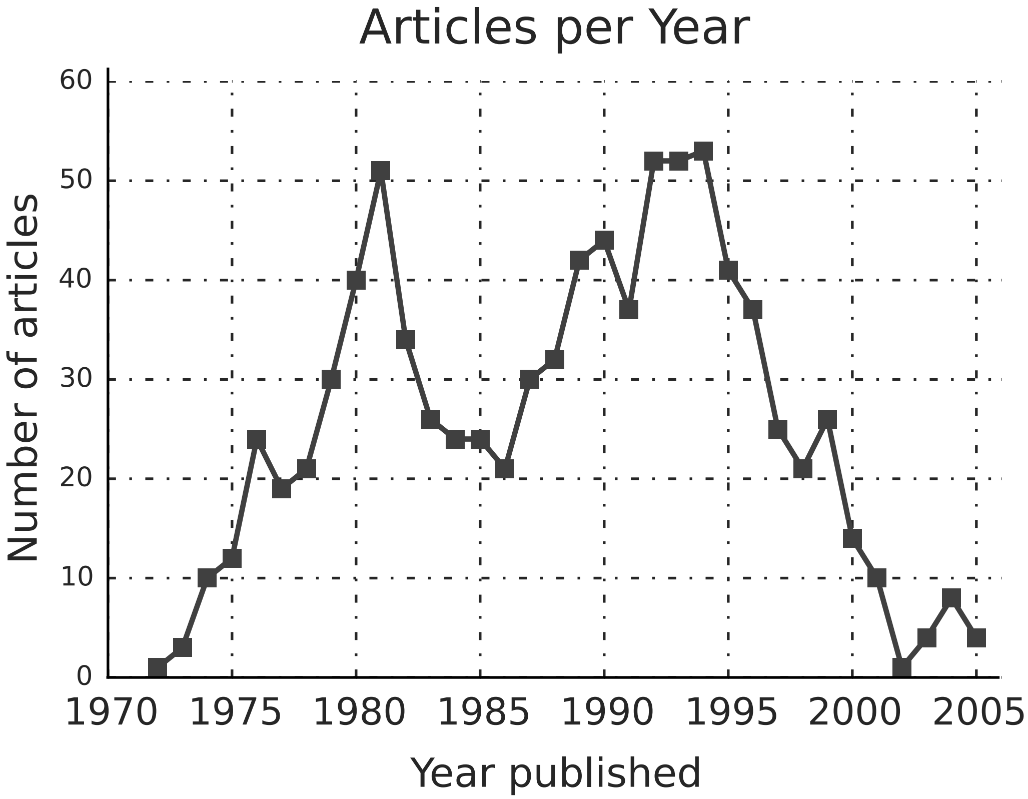

Figure A1. Number of articles per year in the sample (n=1525).

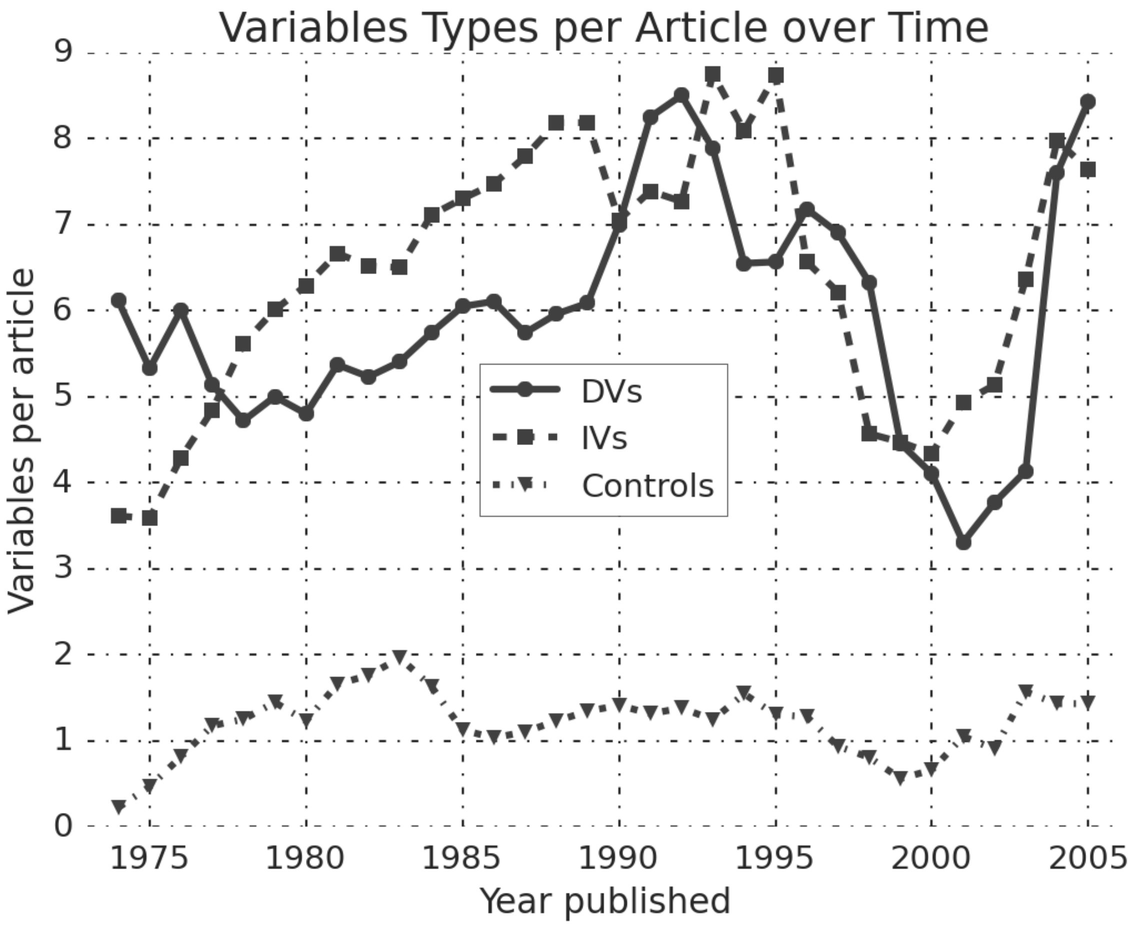

Figure A2. Number of dependent and independent variables per article over time. (3-year moving averages; 1525 articles)

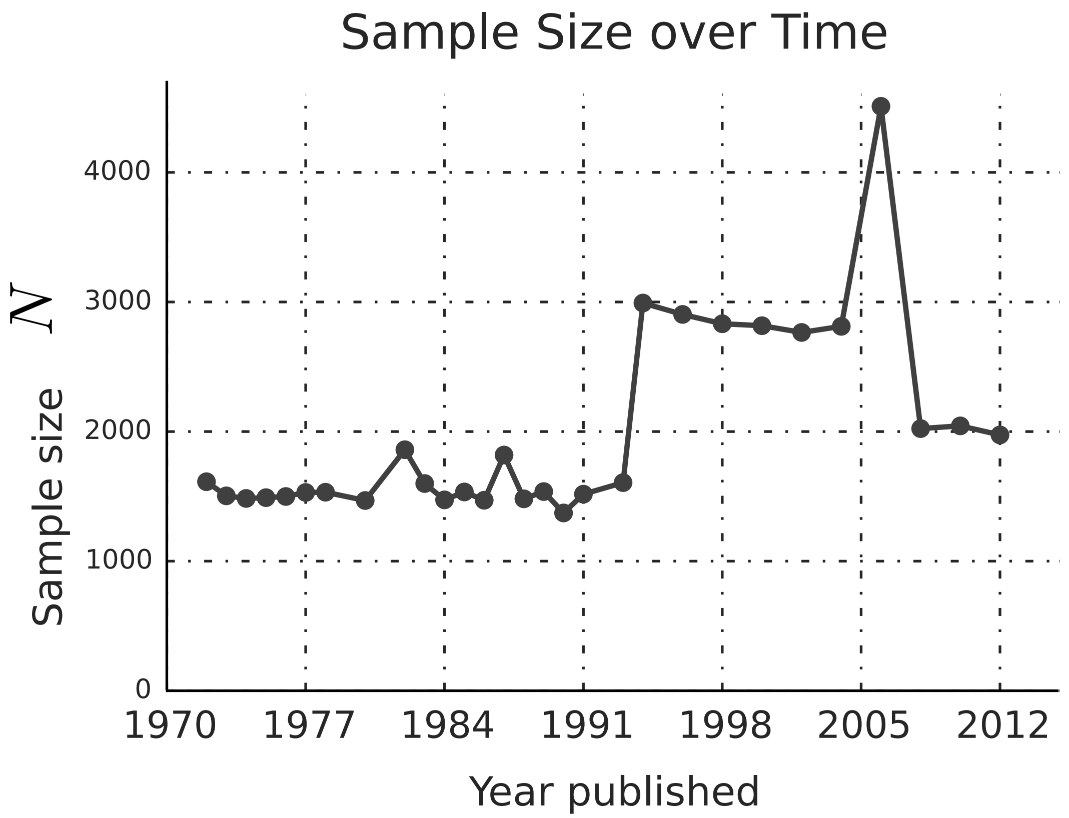

Figure A3. GSS sample size over time. Note that our data does not include articles published after 2005 in order to allow post-publication model estimation.

Approximate Estimation of Original Models

$$\begin{array}{rcccc} Y_{1,t} & = & X_t\beta & + & \epsilon \\ Y_{2,t} & = & X_t\beta & + & \epsilon \\ \vdots & & \vdots & & \vdots \\ Y_{f,t} & = & X_t\beta & + & \epsilon \end{array} \qquad X_t' = \begin{pmatrix} 1 & \cdots & 1 \\ x^{*}_{1,1,t} & \cdots & x^{*}_{1,z,t} \\ \vdots & \ddots & \vdots \\ x^{\circ}_{n,1,t} & \cdots & x^{\circ}_{n,z,t} \end{pmatrix}$$

Examples:

“Race, Sex and Feminist Outlooks” (Ransford and Miller 1983)

$$\begin{array}{lcl} \mathrm{FEHOME}_{1974} & = & X_t\beta + \epsilon \\ \mathrm{FEWORK}_{1974} & = & X_t\beta + \epsilon \\ \mathrm{FEPRES}_{1974} & = & X_t\beta + \epsilon \\ \mathrm{FEPOL}_{1974} & = & X_t\beta + \epsilon \end{array} \qquad X_t' = \begin{pmatrix} 1 & \cdots \\ \mathrm{MAWORK}^{*}_{1,t} & \cdots \\ \mathrm{OCC}^{*}_{1,t} & \cdots \\ \mathrm{EDUC}^{*}_{1,t} & \cdots \\ \mathrm{GOVAID}^{*}_{1,t} & \cdots \\ \mathrm{FINRELA}^{*}_{1,t} & \cdots \\ \mathrm{INCOM16}^{*}_{1,t} & \cdots \\ \mathrm{CLASS}^{*}_{1,t} & \cdots \end{pmatrix}$$

$$t \in \{1974 - 1978\}$$

“Confidence in Science: The Gender Gap” (Fox and Firebaugh 1992)

$$\begin{array}{lcl} \mathrm{CONSCI}_t & = & X_t\beta + \epsilon \\ \mathrm{CONFINAN}_t & = & X_t\beta + \epsilon \\ \mathrm{CONBUS}_t & = & X_t\beta + \epsilon \\ \mathrm{CONCLERG}_t & = & X_t\beta + \epsilon \\ \mathrm{CONEDUC}_t & = & X_t\beta + \epsilon \\ \vdots & & \vdots \\ \mathrm{CONPRESS}_t & = & X_t\beta + \epsilon \end{array} \qquad X_t' = \begin{pmatrix} 1 & \cdots \\ \mathrm{SEX}^{*}_{1,t} & \cdots \\ \mathrm{OCC}^{\circ}_{1,t} & \cdots \\ \mathrm{EDUC}^{\circ}_{1,t} & \cdots \\ \mathrm{PRESTIGE}^{\circ}_{1,t} & \cdots \\ \mathrm{WRKSTAT}^{\circ}_{1,t} & \cdots \\ \vdots & \ddots \\ \mathrm{RELIG}^{\circ}_{1,t} & \cdots \end{pmatrix}$$

$$t \in \{1973 - 1989\}$$

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