The Enduring Value of Social Science Research: The Use and Reuse of Primary Research Data[1]

Amy M. Pienta, George Alter, Jared Lyle

Inter-university Consortium for Political and Social Research, Institute for Social Research, University of Michigan

The goal of this paper is to examine the extent to which social science research data are shared and assess whether data sharing affects research productivity tied to the research data themselves. We construct a database from administrative records containing information about thousands of social science studies that have been conducted over the last 40 years. Included in the database are descriptions of social science data collections funded by the National Science Foundation and the National Institutes of Health. A survey of the principal investigators of a subset of these social science awards was also conducted. We report that very few social science data collections are preserved and disseminated by an archive or institutional repository. Informal sharing of data in the social sciences is much more common. The main analysis examines publication metrics that can be tied to the research data collected with NSF and NIH funding – total publications, primary publications (including PI), and secondary publications (non-research team). Multivariate models of count of publications suggest that data sharing, especially sharing data through an archive, leads to many more times the publications than not sharing data. This finding is robust even when the models are adjusted for PI characteristics, grant award features, and institutional characteristics.

[1] We would like to acknowledge the National Library of Medicine (R01 LM009765). We also thank Darrell Donakowski, Myron Gutmann, Felicia LeClere, JoAnne O'Rourke, James McNally, Russell Hathaway, Kristine Witkowski, Kelly Zidar, Tannaz Sabet, Lisa Quist, and Robert Melendez for their contributions to the LEADS database at ICPSR. The creation of the database was also supported by the following research projects at ICPSR: P01 HD045753, U24 HD048404, and P30 AG004590. Email Address of Corresponding Author: apienta@umich.edu

Introduction

Federal funding for scientific research has always been a highly competitive endeavor with a small proportion of research grant submissions receiving awards from the National Institutes of Health (NIH) each year. The impact, or success, of a funded research project is measured, partly, by the research productivity of the PI and his or her research team who publish findings from primary data collection activities. Increasingly, NIH and the National Science Foundation (NSF) have become interested in data sharing as a means of supporting the scientific process and ensuring the highest return on competitive investments. However, there has been little investigation of research productivity that extends beyond the primary of use of data to test hypotheses outlined in the grant application that led to the data collection activity. We proposed to redress this gap by examining research productivity of research data, both as it applies to the original research team and reuse by members outside the research team.

This research question is particularly salient for the social sciences because several social science disciplines were among the earliest to organize efforts to share research data – so the avenues for sharing data have been fairly well known, especially in the social science disciplines of political science, sociology and economics. Social science research occurs in other social and behavioral disciplines, as well. So, there is tremendous heterogeneity in data sharing in the social sciences. As a result, despite the best efforts of social science data archives in the United States, many social science studies do not reside in a permanent archive (Pienta et al. 2008).

The largest share of social science research is conducted with federal support. The National Science Foundation (NSF) and the National Institutes of Health (NIH) historically have supported a significant share of social science data collections and the trend continues today (Alpert 1955; Alpert 1960; Kalberer 1992). This paper focuses on analyzing information from grant awards made by NSF and NIH, making it possible to enumerate the bulk of the major social science data collections that exist today. Also, NSF and NIH keep electronic records about grant awardees that can be and have been culled into a single database useful for understanding the scope and breadth of social science research that has produced research data. Thus, this research topic is both timely and practical.

Background

Data sharing has been an important topic of debate in the social sciences for more than twenty years, initially spurred by a series of National Research Council Reports (1985) and more recently the publication of the National Institutes of Health Statement on Sharing Research Data in February 2003 (NIH 2003). Despite this formal written statement from NIH and a similar one from the National Science Foundation (NSF-SBE n.d.) that give official support for the long held expectations placed on grantees to share their research data, little is known about the extent to which data collected with support from NIH or NSF has been shared with other researchers. The limited work done suggests considerable variability in the extent to which researchers' share and archive research data. Our LEADS database of NIH and NSF grant information will fill this gap in knowledge and create a research database for answering these questions.

NIH's policy is designed to encourage data sharing with the goal of advancing science. The benefits of sharing data have been widely discussed and understood by researchers for years. An important part of Kuhn's (1970) scientific paradigm is the replication and confirmation of results. Sharing data is at the core of direct replication (Anderson et al. 2005; Kuhn 1970; Freese 2006). The foundation of the scientific process is that research should build on previous work, where applicable, and data sharing makes this possible (Bailar 2003; Louis, Jones & Campbell 2002). The argument has been made, and there is some evidence to support it, that sharing data and allowing for replication makes one’s work more likely to be taken seriously and cited more frequently (King et al. 1995). In fact, Glenditsch, Petter, Metelits, and Strand (2003: 92) find that authors who make data from their articles available are cited twice as frequently as articles with “no data but otherwise equivalent credentials, including degree of formalization.”

Additionally, the nature of large datasets virtually guarantees that a single researcher or group of researchers will not be able to use the dataset to its full potential for a single project. It may be the case that those who collect the data are not the best at analyzing them beyond basic descriptive analyses (Bailar 2003). Sharing data in this way ensures that resources spent on data collection are put to the best use possible and the public benefit is enhanced.

Finally, the use of secondary data is crucial in the education of undergraduate and graduate students (Fienberg 1994; King 2006). It is not feasible for students in a semester-long course to collect and analyze data on a large scale. Using datasets that have been archived and shared allows students to experience science firsthand. Instructors can use the metadata accompanying shared data to teach students about “good science” and the results obtained from even simple analyses to illustrate the use of evidence (data) in support of arguments (Sobal 1981).

Policies about Data Sharing

Most institutes and organizations that finance research, especially data collection, have a policy about sharing data once the initial project is completed. The National Institutes of Health (NIH 2003) and National Science Foundation (NSF-SBE n.d.), for example, require a clearly detailed plan about data sharing as part of research proposals submitted for review. Plans must cover how and where materials will be stored, how access will be given to other researchers, and any precautions that will be taken to protect confidentiality when the data is made public. These requirements are not, however, evaluated in the review process nor are there formal penalties for non-compliance after the award. Most professional organizations also include a statement in their "best practice" or ethics guidelines that addresses the issue that research reports should be detailed enough to allow for replication and that researchers should also make available their data and assistance in these replication attempts when the requests are made (e.g., American Sociological Association, American Psychological Association, American Association for Public Opinion Research).

In addition to such general statements that data collected with public funds must be shared with other researchers and that individuals should be willing to assist others replicating their work, some fields, such as Economics, have taken steps to make the data sharing policy more concrete. In an attempt to allow for direct replications as well as full-study replications, the American Economic Review and other major economics journals have instituted the practice that any article to be published must be accompanied by the data, programs used to run the analyses, and clear, sufficient details about the procedures prior to publication (Freese 2006; Anderson et al. 2005). The requirement to include not only the data but also statistical code written to perform analyses requires that individual researchers thoroughly and carefully document decisions made during the analysis stages of the project and allows other researchers to more easily use these as starting points for their own work. This has led to an increased use and citation of work that has been published in journals where this type of information is required (Anderson et al. 2005; Glenditsch et al. 2003).

Sharing Social Science Data

Data are currently shared in many different ways ranging from formal archives to informal self-dissemination. Data are often stored and disseminated through established data archives. These data generally reach a larger part of the scientific community. Also, data in formal archives typically include information (metadata) about the data collection process as well as any missing data imputations, weighting, and other data enhancements. These archiving institution have written polices and explicit practices to ensure long-term access to the digital assets that they hold that include replication copies stored off-site and a commitment to the migration of data storage formats. These are the characteristics that define data archives.

Another tier of data archives have more narrowly focused collections around a particular substantive theme such as the Association of Religion Data Archives (http://www.thearda.com). The data in these kinds of thematic archives are not necessarily unique, though some of their holdings are, but the overlap between archives makes data available to broader audiences than might be captured by a single archive. The ARDA, for instance, has a broader non-scientific audience who are interested in analysis and reports as well as the micro-data files for reanalysis. These archives expend resources on the usability of the collection for the present day and make some kind of a commitment to long-term access through migration and back-ups.

Some data archives are designed solely to support the scientific notion of replication. Journal-based systems of sharing data have become popular in Economics and other fields as a way of encouraging replication of results (Anderson et al. 2005; Glenditsch et al. 2003). The longevity of these collections is sometimes more tenuous than the formal archives, particularly if the sustainability of their archival model relies on a single funding source.

Some examples of less formal approaches include authors who acknowledge they will make their data available upon request or who distribute information or data through a website. Researchers often keep these sites up to date with information about findings from the study, and publication lists in addition to data files and metadata. These sites are limited to those who know about the study by name or for whom the website has shown up in an internet search (see also Berns, Bond & Manning 1996). Typically, the commitment to preserving the content lasts only as long as the individual has resources available.

The Reluctance of Researchers to Archive Data

The time and effort required to produce data products that are useable by others in the scientific community is substantial. This extra effort is seen by many as a barrier to sharing data (Birnholz & Bietz 2003; Stanley & Stanley 1988). In addition to the actual data, information must be added to assist secondary users in identifying whether the data would be of value to them and in the analysis and interpretation of results. Such metadata includes complete descriptions of all stages of the data collection process (sampling, mode of data collection, refusal conversion techniques, etc.) as well as details about survey question wording, skip patterns and universe statements, and post-data processing. All of these factors allow subsequent researchers to judge the quality of the data they are receiving and whether it is adequate for their research agenda. Therefore, substantial effort is required of those sharing data, while the benefits accrue to the secondary user.

Another significant barrier in the sharing of data is the risk of breaching the confidentiality of respondents and the potential for the identification of respondents (Bailar 2003). The issue of protecting confidentiality has become more salient as studies collect information about social context, which may include census tract or block group identification to allow researchers to link the data collected with information about the context. Not only are data about social and community contexts being collected and included in datasets but also global positioning coordinates and information about multiple members of a household, all of which could make identification of any single individual easier. Additional information about biomarkers and longitudinal follow up are also hallmarks of new data collection efforts. Both methodological innovations make it more difficult for Institutional Review Boards to allow for the wide redistribution of data.

Other reasons individuals give for withholding data include wanting to protect their or their students' ability to publish from the data as well as the extra effort involved in preparing data for sharing (Louis et al. 2002). Retaining the ability to publish from one's data seems to be a significant concern among scientists, both for fear of others "scooping" the story and that others will find mistakes in their attempt to replicate results (Anderson et al. 2005; Bailar 2003; Freese 2006.

Current publication and academic promotion practices act as another barrier to sharing data – or, put another way, those who "hoard" their data are likely to be rewarded more than those who "share". There are often few, if any, rewards to sharing data, especially given the expense in terms of time and effort required to prepare clean, detailed data and metadata files. Researchers are not typically rewarded for such behavior, particularly if the time spent on data sharing tasks infringes on one's ability to prepare additional manuscripts for publication. Academic culture does not support the scientific norm of replication and sharing with tangible rewards. (Anderson et al. 2005; Berns et al. 1996). As an example, in discussing the notion that researchers might share not only data but also analytic/statistical code, Freese (2006:11) notes that a typical reaction to a "more social replication policy would be to expend less effort writing code, articulating a surprisingly adamant aversion to having [one's] work contribute to others' research unless accompanied by clear and complete assurance in advance that they would be credited copiously for any such contribution." It is unlikely that attitudes about data sharing will change without strong leadership and examples set by senior scientists and the commitment of scientific institutions such as universities and professional societies who facilitate and enforce such sharing (Berns et al. 1996).

Extending Research Productivity to include Data Reuse

Research productivity is often thought of as something that scientists accomplish by publishing their research discoveries. The second part of research productivity is not how many times your ideas are published, but also how often the idea is cited in the work of others (Matson, Gouvier, Manikam 1989). This is an analysis of citation counts of a scientist's publications – how widely cited their publications are. Thus, the impact of a scientist's scholarship is derived directly from their own published work.   However, there has been movement in the scientific academy to recognize the importance and value of research data. We consider the possibility that research data may have enduring value on scientific progress as scientists use and reuse research data to draw new analysis and conclusions. This idea is rooted in the idea of a data life cycle – where research data can often have use beyond its original designed purpose (Jacobs and Humphrey 2004).   This is not farfetched given research productivity measures have also been used to assess institutional productivity across universities (Toutkoushian, Porter, Danielson, and Hollis 2003). Here, we consider the research productivity attached to research data collected by a scientist with federal funding.

In summary, while the social sciences share in the normative expectation that research data must be shared to foster replication and reanalysis, there is little to suggest that it is a wide spread practice. Federal institutions and professional organizations underscore these normative expectations with implicit and explicit sharing policies. The advantages of sharing data with the research community are large and cumulative. Yet, with the exception of leading journals in Economics, there are few cases in which these normative statements are coupled with penalties or incentives to reinforce them. The institutional, financial, and career barriers to data sharing are substantial as noted. What remains an open empirical question is the extent of data sharing across social science disciplines and the value this has for the social sciences.

Methods

To address this question we construct a database of research projects. The LEADS database is comprised of social and behavioral science awards made by NSF and NIH. From the National Science Foundation online grants database (http://www.nsf.gov/awardsearch/), we include in our study research grant awards from three NSF organizations that matched prominent search terms relating to the social sciences[2]. We further restrict this set of awards to awards that include descriptions of research activity that (1) relates to the social and /or behavioral sciences and (2) likely includes original (or primary) data collection (including assembly of a new database from existing or archival sources). From the National Institutes of Health online CRISP (Computer Retrieval of Information on Scientific Projects) database (http://crisp.cit.nih.gov/)[3], we include extramural research grant awards from the top ten NIH institutes engaging in social and behavioral

[2] We determined this set of organizations in a pretest. The NSF organizations we include in the sample are: Directorate for Computer & Information Science & Engineering, Office of Polar Programs, and Directorate for Social, Behavioral & Economic Sciences. Search terms used to select possible awards from these NSF organizations for inclusion in LEADS included: SOC*, POLIT*, and/or STAT*.

[3] CRISP has since been replaced by the RePORT Expenditures and Results (RePORTER) query tool (http://projectreporter.nih.gov/reporter.cfm), although all NIH information described in this paper was collected using the CRISP database.

research[4]. In additional to screening these awards for social and behavior science content, these awards also were restricted to the collection of original quantitative data. This strategy differs from the NSF award review in that strictly qualitative studies were not identified as such and excluded from LEADS.

Of the 235,953 eligible NSF and NIH awards in the LEADS database, 12,464 matched our initial screening criteria (i.e., social/behavior science & collected research data). We select awards from 1985-2001 (n=7,040). We selected this range of years because we wanted to inquire about completed research that could have led to publications and data archiving. But, we did not want to select awards that were completed so long ago that recall of information about the publications related to the award would be unreasonable. From this set of awards, we determined there are 4,883 unique PIs. We attempted to invite all 4,883 of these PIs[5] to complete a web survey[6].

The PI survey consisted of consisted of questions about research data collected, various methods for sharing research data, attitudes about data sharing and demographic information. PIs were also asked about publications tied to the research project including information about their own publications, research team publications, and publications outside the research team. We received 1,217 responses (24.9% response rate). For the analytic sample we select PIs and their research data if (1) they confirm they collected research data (86.6% of the responses), (2) they did not collect data for a dissertation award (n=33), or (3) they were missing data on the dependent variable.

[4] We determined this set of institutes in a pretest. The NIH institutes we include in the sample are: Agency for Healthcare Research and Quality, National Institute of Child Health and Human Development, National Institute of Mental Health, National Institute of Nursing Research, National Institute on Aging, National Institute on Alcohol Abuse and Alcoholism, National Institute on Drug Abuse

[5] While we attempted to e-mail survey invitations to all investigators, we were unable to reach at least 1,632 PIs after one or more attempt. This was due to PIs no longer living, non-public e-mail addresses (i.e., e-mail addresses not readily discoverable on the Web), and non-working e-mail addresses (i.e., 'bounced' e-mails).

[6] PIs with multiple awards were asked about just one of their awards, as selected randomly.

Publication Measures – Research productivity is typically assessed by either or both citation and publication analysis. The outcome measures used in this analysis are various measures of publication counts. Publication counts are based on self-reported information provided by PIs of the research grant awards at NSF and NIH. PIs are asked to report number of publications related to the data they collected including estimates for: own publications, publications of the research team, extant publications not related to the research team and the number of publications (in each of the three previous categories) that include students. We include in this analysis count of publications where the PI is one of the authors (range 0 – 100). This is the first measure of primary publications. A second measure of primary publications is created that also includes counts of publications where the PI may or may not be an author, but at least one member of the research team is an author (range 0-350). Secondary publications are publications where none of the original research team (PI, co-investigators, students or other researchers) is a co-author of the publication (range 0-700). This measure indicates the extent of reuse (or secondary use) of research data beyond its original collection purpose. Next, total publication count is constructed by adding counts of all primary publications with counts of secondary publications (range=0-713). Finally, the number of publications where a student was author or co-author is defined (range 0-160).

The main independent variable used in the analysis is data sharing status. PIs are asked whether they share the research data from the award (1) formally through a data archive (or institutional repository), (2) informally, not through a data archive (including shared upon request, personal website, departmental website), or (3) not shared.

To ensure that data sharing is not “standing in” for other known predictors of productivity, we include covariates describing characteristics of the individuals who collect the data, the award mechanism used to fund the data, and the institutional home of the original data collection.

Research productivity has been linked to departmental prestige (Long 1978), age (Levan and Stephan 1991) and gender (Penas and Willet 2006) among other factors. We begin by describing PI characteristics we are able to measure.

PI Characteristics – We expect that characteristics of the PIs themselves will be associated with both data sharing status and various publication counts. Some researchers have more time for archiving and publishing whereas others may be more likely to engage in training and service. We attempt to control for this by including various social and demographic characteristics of the PIs in the models. The gender of the PI is male (=1) or female (=0)[7]. The self-reported race/ethnicity of the PI is defined as white (=1) versus non-white (=0). Age (in years) at time of initial award is calculated by subtracting year of birth from year at start of initial award (range 27-75). Self-reported faculty status/rank at time of initial award is defined as senior (tenured faculty), junior (untenured faculty), and non-faculty (including students, postdocs, research staff). Self-reported discipline is classified from an open-ended question and collapsed into the following categories: (1) health sciences (nursing, medicine, public health) and psychology, (2) core social science (political science, sociology, and economics), and (3) other social science-related discipline (anthropology, film, communications)[8]. Finally, the number of federal grants awarded throughout one's career is defined as number of self-reported federal research grants (range 1-100).

Institutional Characteristics – Next, we construct a set of measures about the institutions awarded the research grant – the institution of the PI at time of initial award. First, we use the

[7] A small number of cases were missing data about self-reported gender. For these, we used a constructed measure of gender where gender was assigned using a tool that analyzed the common first names and estimated gender. This measure was constructed for the full database, not just the PIs who were part of the survey. The tool constructed names based on the top 100 names from the Social Security Baby Names index for 1930-1975 (in 5-year increments). See: http://www.ssa.gov/OACT/babynames/.

[8] We combine health sciences and psychology because neither of the disciplines typically shares data. We combine the core social science disciplines (sociology, political science, and economics) because the rates of data sharing were similar across the three disciples. The other category shares data at rates midway between the other two categories.

Carnegie Classification[9] to differentiate research institutions from non-research institutions. Research institutions include research universities, doctoral granting universities, and medical schools/centers). Non-research institutions include 2- and 4- year colleges, Master's colleges and universities, professional institutions and tribal colleges. Other institutions not classified under Carnegie are divided into private research organizations[10] and other non-Carnegie institutions[11]. The other institutional characteristic defined is the region where the institution is located. This is defined into the following categories: northeast, south, midwest and west.

Grant Award Characteristics – Two measures of the grant award are defined. First, we differentiate awards made by the National Science Foundation from the National Institutes of Health. NSF has had in place a data sharing policy for a longer time and it is expected that data will be shared and archived more frequently when funded by NSF. The other award measure is the duration of the award, measured in years (range =0-8 years). This is included to capture size of award.

Analysis Plan – Descriptive statistics are calculated using univariate and bivariate statistics. Because the outcome measures are publication counts, Poisson regression models are considered. Overdispersion led us to the choice to estimate negative binomial regression models of publication counts for longitudinal data (offset by the amount of time between initial award and the survey). We estimate two sets of models for each outcome. First, we estimate models that include only a three category data sharing status measure. The second set of models adds the various PI, institution, and award characteristics. We do not include any covariates in the final models shown (model 2) that were not statistically significant across all outcomes. The hierarchical set of models

[9] Carnegie Classification information was based on the "Classifications Data File" (Last update: April 12, 2009). See: http://classifications.carnegiefoundation.org/resources/

[10] Private Research Organizations include Non-academic institutions with a research mission (e.g. American Bar Foundation)

[11] These include: Non-US institutions (e.g. University of Oxford) and awards made to individuals

(model 1 and model 2) allows us to understand the extent to which differences by data sharing status might be attributed to other characteristics of PIs, institutions and the awards.

Results

Descriptive sample characteristics are presented in Table 1. The sample of PIs is fairly evenly dived between males (51.9 percent) and females (48.1%). The majority of the sample is white (86.8 percent) and tenured (54.3 percent). Only 20 percent of the sample of PIs is non-faculty. The mean number of Federal grants the PIs have been awarded throughout their careers is 6.2. The majority of PIs come from either the psychological or health sciences (62.5%). Just over a quarter of the sample are PIs in the core social science disciplines (sociology, economics and political science).

[Insert table 1 about here]

The sample of PIs comes from all four major regions of the U.S. The largest numbers of grant awards are made to institutions located in the northeast (36 percent) and the fewest number of grant awards are made to institutions located in the west (18.7 percent). The vast majority of PIs of the research grant awards are working at institutions classified by the Carnegie Classification as research institutions (78.8 percent). The second largest institution type represented in the PI survey is private research organizations (12.3 percent). Few awards were made to non-research institutions and other types of organizations not reflected in the Carnegie Classification (6.5 percent and 2.5 percent respectively). Only 27.3 percent of the awards in our sample come from the National Science Foundation, with the majority coming from the National Institutes of Health (72.7%). The mean length of award is 3.1 years. Few awards produce research data that are shared formally –

either in a data archive or institutional repository (11.5%). Of the rest, half the data from the awards are shared informally, not in an archive (44.6 percent), and half are not being shared beyond the research team (43.9%).

Turning to Table 2, we next examine how various characteristics of the PIs, institutions and grant awards are related to data sharing status. Women are more likely to archive data than men (12.0 and 8.1 percent respectively; chi-squared is statistically significant). We see that PIs who are senior faculty are more likely to archive data as well (12.0 percent) – nearly twice as often as junior faculty (7.1 percent) and non-faculty (8.6 percent). There are strong disciplinary differences as well. The core social science disciplines archive data at the highest rate (27 percent). Psychologists and health scientist archive data least often (4.6 percent). PIs at institutions located in the south are least likely to archive data (8.5 percent). The Carnegie Classification of the institution awarded a research grant to collect data is not associated with data sharing status. Data funded by NSF research grant awards are nearly three times more likely to be archived than data funded by NIH.

[Table 2 about here]

Table 3 shows the distribution of various publication counts for the full sample and by data sharing status. For example, PIs write (or contribute to) a median of 4 publications. However, PIs who archive data formally write a median of 6 publications – compared to PIs who do not share data (3 primary publications). Research teams are also more productive when they archive the data. The median number of research team publications is 8 when data are archived compared to 3 when data are not shared outside the research team.

[Table 3 about here]

Large numbers of research data contribute to no secondary publications beyond the PI and research team. Thus, across all categories, the median number of secondary publications is 0. For this outcome we examine the mean. A research grant award produces 2 secondary (non research team) publications on average. However, when data are archived 4 secondary publications are reported on average. We turn to the total number of publications next. A research grant award produces a median of 5 total publications. However, when data are archived a research grant award leads to a median of10 publications. When data are shared informally a research grant is linked to a median of 7 publications. And, when data are not shared outside the research team, the research data lead to only a median of 4 publications overall. The same pattern is found for publications with student authorship as well.

Multivariate results are presented in Tables 4.[12] Count of total publications is related to data sharing status. Both archiving data and sharing data informally are related positively to count of total publications (b=1.094 and b=1.020 respectively). Both associations are statistically significant (p<.01). This can be interpreted (by taking the exponential of the log-odds) as archiving data leading to 2.98 times more publications than not sharing data. When data are shared informally (compared to not shared at all), 2.77 times the number of publications is produced.

[Table 4 about here]

Turning to model 2, additional covariates are added to the model to account for potential differences in PIs, institutions, and the grant awards. The coefficients for archiving data and informally sharing data are positively associated with number of total publications count in comparison to not sharing data at all. These two coefficients are smaller than in model 1, but still statistically significant. This can be interpreted (by taking the exponential of the log-odds) as archiving data leads to 2.42 times more publications than not sharing data. Informally sharing data leads to 2.31 times more publications than not sharing data at all. Thus, the effect of data sharing,

[12] Dispersion differs from 0 across all outcomes and models supporting negative binomial regression models. Log-likelihood estimates are presented in Tables 4 & 5, standard errors appear in parentheses.

formally or informally, is not explained by differences in the PI themselves, the awards or the institutions that were given the awards to conduct the research. Research productivity benefits clearly from data sharing, particularly archiving data.

Other coefficients in the model demonstrate that being older at the time of award is associated with increasing log-odds of total publications. Being older at time of award may translate into a measure of writing and publishing experience – and in turn older PIs may have a publishing advantage that is not explaining by other factors. One of the surprising results is that faculty status (senior, junior, and non-faculty) at time of award was not statistically significant in the model. This covariate (and gender, race and number of federal grants) are not included in model 2. The other PI characteristic that affects total publications is PI's discipline. Compared to data collected by core social scientists, data collected by health scientists and psychologists have lower log-odds of leading to overall publications (b=-.254).

Only one measure of the institutional climate surrounding the award that produced data is retained in model 2. Carnegie Classification is associated with total publication count. Compared to data collected at research universities, data collected at non-research institutions reduce the log-odds of overall publications (b=-.685). Data collected at other non-Carnegie classified institutions (but not private research organizations that were classified separately), compared to data collected at Carnegie research universities, are actually associated with increased log-odds of publications (b=.230). Finally, the greater the length of the initial award period the greater the log-odds of publication (b=.199).

The next set of models examines the number of secondary publications. Secondary publications are publications by researchers outside the research team. We find that secondary publications are also related to data sharing status. Both archiving data and sharing data informally are positively related (increase the log-odds) of secondary publications (b=2.515 and b=2.375 respectively). Both associations are statistically significant (p<.01). This can be interpreted (by taking the exponential of the log-odds) as archiving data leads 12.37 times more publications than not sharing data. When data are shared informally (compared to not shared at all), 10.75 times the number of publications are produced.

Turning to model 2, additional covariates are added to the first model to account for potential differences in PIs, institutions, and the grant awards. The coefficient for archiving data is positively associated with secondary publication count in comparison to not sharing data at all, but is smaller than in model 1 (b=1.919 in model 2 compared to b=2.515 in model 1). This can be interpreted (by taking the exponential of the log-odds) as archiving data leads to 6.81 times more secondary publications than not sharing data. Both archiving and informal sharing are positive and statistically significant in model 2 (p<.01). Informal data sharing leads to 4.78 times more secondary publications than not sharing data. Thus, the effect of data sharing, formally or informally, is not explained by differences in the PI themselves, the awards or the institutions that were given the awards to conduct the research. Data reuse that leads to research productivity is tied closely to data sharing, particularly archiving data.

The remaining covariates in model 2 have similar relationships with secondary publications as total publications. A few notable differences emerge. Other social science disciplines differ from the core social science disciplines in that data collected by other social scientists have lower log-odds of secondary publications. The other difference is that research data collected by private research organizations are related to greater log-odds of secondary publication compared to research universities (b=1.230). This reinforces the idea that scientists at private research organizations collect data to be shared externally.

Multivariate results are also presented in Tables 5.[13] The first set of models shows that primary publications (PI included as an author) are also related to data sharing status. Archiving data is positively related to the log-odds of primary PI publications (b=.620). Informal data sharing is positively related to the log-odds of primary PI publications (b=.743). Both associations are statistically significant (p<.01). This can be interpreted (by taking the exponential of the log-odds) as archiving data leading to nearly 2 times more publications than not sharing data. Adding the additional covariates in model 2 does not explain the data sharing effects. In the last set of models we saw private research organizations (PROs) produce data that lead to greater numbers of secondary publications. Here, in model 2, we see that PROs produce data that lead to lower log-odds of primary PI publications compared to research universities (b=-.216). Also in this model, we see that NIH data increase the log-odds of primary publications compared to NSF data.

[Table 5 about here]

Much like the other publication metrics, the number of publications including students is related to data sharing status. Archiving (b=.700) and sharing data informally (b=.763) increase the log-odds of publications including students in comparison to not sharing data. Adding the additional covariates in model 2 does not explain data sharing differences.

Conclusions

The LEADS database contains valuable information about a wide range of social science research data collected with support from the National Science Foundation and the National Institutes of Health. NSF and NIH awards typically lead to some of the largest investigator-initiated

[13] Dispersion differs from 0 across all outcomes and models supporting negative binomial regression models. Log-likelihood estimates are presented in Tables 4 & 5, standard errors appear in parentheses.

research activities in the U.S. and both institutions have had longstanding expectations that data collected with public money ought to be made available to the public and/or research community. In the social science research community, more so than in other basic disciplines, there have been longstanding avenues for archiving and sharing data. Even with this advantage, we confirm that the majority of social science data are not archived publicly (88.5%). Informal data sharing, though much more common (44.6%), does not ensure that the scientific information collected with public funding has enduring value beyond its original primary publications.

One of the central questions stemming from this disparity is whether research productivity varies by data sharing. We find strong and consistent evidence that data sharing, both formal and informal, increases research productivity across a wide range of publication metrics. Data archiving, in particular, yields the greatest returns on investment with research productivity (number of publications) being greater when data are archived. Not sharing data, either formally or informally, limits severely the number of publications tied to research data. We hypothesized that some of the data sharing advantage would be explained by PI characteristics and characteristics of the institutions and grant awards. We find that although this is true, large persistent advantages in research productivity accrue when data are shared. Finally, we also include a large number of publication metrics to better understand how data sharing affects primary versus secondary publications. Data sharing is related to all publication metrics, even primary PI publications. However, data sharing has the largest effects on secondary publications as expected. Data archiving, and informal data sharing, generate many more secondary publications and PI and research team exclusive use.

Limitations - It is unclear whether larger numbers of primary publications lead to data sharing or if sharing data leads to more primary publications. While both are plausible, it is likely that the association we observe between data archiving and primary publications reflects the fact that PIs archive data when their research is complete and all primary findings are published. That said, we carefully selected a range of grant awards that would have been completed years ago.

Larger research projects probably lead to more publications and greater likelihood of data sharing. While we have included a measure of grant award duration to get at some of the variability in grant award size, a better measure of the size of the research project is amount of dollars. The largest social science data collections cost more money to collect, are intended for public dissemination, and have more questions that will appeal to a larger number of scientists. Unfortunately, this information was not available for NIH awards.

References

Alpert, Harry. 1955. The Social Sciences and the National Science Foundation. *Proceedings of the American Philosophical Society,* 99(5), Conference on the History, Philosophy, and the Sociology of Science: 332-333.

Alpert, Harry. 1960. The Government's Growing Recognition of Social Science. *Annals of the American Academy of Political and Social Science*, 327, Perspectives on Government and Science: 55-67.

Anderson, Richard G., William H. Greene, B.D. McCullough and H.D. Vinod. 2005. The Role of Data and Program Code Archives in the Future of Economic Research. The Federal Bank of St. Louis Working Paper Series.

Bachrach, Christine. 1984. Contraceptive Practice among American Women, 1973-1982. *Family Planning Perspectives* 16:253-259.

Bailar, John C., III. 2003. The Role of Data Access in Scientific Replication. Paper presented at Access to Research Data: Risks and Opportunities. Committee on National Statistics, National Academy of Sciences.

Berns, Kenneth I., Enriqueta C. Bond, and Frederick J. Manning (eds). 1996. *Resource Sharing in Biomedical Research.* Committee on Resource Sharing in Biomedical Research, Division of Health Sciences Policy, Institute of Medicine. Washington, D.C.: National Academy Press.

Fienberg, Stephen E. (1994). Sharing Statistical Data in the Biomedical and Health Sciences: Ethical, Institutional, Legal, and Professional Dimensions. *Annual Review of Public Health* 15:1-18.

Freese, Jeremy. 2006. Replication Standards for Quantitative Social Science: Why Not Sociology? Unpublished manuscript, University of Wisconsin-Madison.

Glenditsch, Nils Petter, Claire Metelits, and Havard Strand. 2003. Posting Your Data: Will You be Scooped or Will You Be Famous? *International Studies Perspectives* 4(1):89-95.

Jacobs, James A., and Charles Humphrey. (2004). “Preserving Research Data.” *Communications of the ACM. 47*(9): 27–29.

Kalberer, Jr., John T., 1992. When Social Science Research Competes with Biomedical Research. *Medical Anthropology Quarterly*, New Series, 6(4):391-394.

King, Gary. 2006. Publication Publication. *Political Science & Politics*, 39(1):119-25.

King, Gary, Paul S. Herrnson, Kenneth J. Meier, M.J. Peterson, Walter J. Stone, Paul M. Sniderman, et al. 1995. Verification/Replication. *PS: Political Science and Politics* 28(3):443-499.

Kuhn, Thomas. 1970. *The Structure of Scientific Revolutions*. Chicago: University of Chicago Press.

Levan, Sharon G. and Paula E. Stephan. 1991. Research Productivity Over the Life Cycle: Evidence for Academic Scientists, *American Economic Review 81*, 1, 114-132.

Long, J. Scott. 1978. Productivity and Academic Position in the Scientific Career. *American Sociological Review 43,* 6, 889-908.

Louis, Karen Seashore, Lisa M. Jones, and Eric G. Campbell. 2002. Sharing in Science. *American Scientist* 90(4): 304-307.

Matson, Johnny L., William Drew Gouvier, and Ramasamy Manikam. 1989. Publication Counts and Scholastic productivity: Comment on Howard, Cole and Maxwell. *American Psychologist*, 737-739.

National Institutes of Health (NIH). 2003. *Final Statement on Sharing Research Data*. February 26, 2003. Retrieved September 6, 2006 from http://grants.nih.gov/grants/policy/data_sharing/

National Research Council. 1985. *Sharing Research Data*. Stephen E. Fienberg, Margaret E. Martin, and Miron L. Straf, Eds. Committee on National Statistics. Washington, D.C.: National Academy Press.

National Science Foundation Directorate for Social, Behavioral, and Economic Sciences (NSF-SBE). (n.d.) *Data Archiving Policy*. Retrieved August 21, 2006 from www.nsf.gov/sbe/ses/common.

Pienta, Amy, Myron Gutmann, Lynette Hoelter, Jared Lyle, and Darrell Donakowski. 2008. The LEADS Database at ICPSR: Identifying Important "At Risk" Social Science Data. Roundtable paper presentation at the American Sociological Association Annual Meeting 2008, Boston, MA.

Robbin, Alice. 2001. The Loss of Personal Privacy and Its Consequences for Social Research. *Journal of Government Information* 28(5): 493-527.

Sobal, Jeff. 1981. Teaching with Secondary Data. *Teaching Sociology*. 8(2): 149-170.

Stanley, Barbara and Michael Stanley. 1988. Data Sharing: The Primary Researcher's Perspective. *Law and Human Behavior* 12(2): 173-180.

Toutkoushian, R.K., Porter, S.R., Danielson, C., & Hollis, P.R. (2003). Using publication counts to measure an institution's research productivity. *Research in Higher Education*, 44, 121–148.

Table 1. Descriptive Sample Characteristics (n=930)

	Total	Range
PI Characteristics		
Female (%)	48.1	
Primary Publications (w/ PI)		
White (%)	86.8	
Primary Publications (w/any Research Team Member)	43.4	27-75
Faculty Status @ Award - Senior (%)	54.3	
Faculty Status @ Award - Junior (%)	25.7	
Faculty Status @ Award) - Non-Fac (%)	20.0	
Discipline - Core Social Science	25.5	
Discipline - Psychology & Health	62.5	
Disciple - Other	12.0	
# Fed Grants in Lifetime (mean)	6.2	1-100
Institutional Characteristics		
Region - Northeast (%)	36.0	
Region - Midwest (%)	23.7	
Region - South (%)	21.6	
Region - West (%)	18.7	
Carnegie-Research (%)	78.7	
Carnegie-Non Research (%)	6.5	
Carnegie-Uncl, PRO (%)	12.4	
Carnegie-Uncl, Other (%)	2.5	
Grant Characteristics		
NSF Award (%)	27.3	
Duration of Initial Award, Years	3.1	0-8
Data Sharing Status		
Shared Formally, Archived	11.5	
Shared Informally, Not Archived	44.6	
Not Shared	43.9	

Table 2. Bivariate Relationships: Data sharing status by PI Characteristics, Institutional Characteristics, and Grant Award Characteristics

	Shared Formally, Archived (n=111)	Shared Informally, Not Archived (n=415)	Not Shared (n=409)	p-value
PI Characteristics				
Female (%)	15.1	42.2	42.7	***
Male	7.6	47.2	45.2	
White (%)	12.0	45.5	42.5	*
Nonwhite	8.1	39.0	52.9	
Age @ award (mean)	44.3	43.4	43.1	
Faculty Status @ Award - Senior (%)	14.7	45.7	39.6	***
Faculty Status @ Award - Junior (%)	7.1	48.5	44.4	
Faculty Status @ Award - Non-Fac (%)	8.6	36.6	54.8	
Discipline - Core Social Science	27.0	48.5	24.5	***
Discipline - Psychology & Health	4.7	42.9	52.5	
Disciple - Other	14.3	45.5	40.2	
# Fed Grants in Lifetime (mean)	7.3	6.3	5.8	
Institutional Characteristics				
Region - Northeast (%)	29.7	39.3	34.5	*
Region - Midwest (%)	15.5	43.6	40.9	
Region - South (%)	8.5	45.3	46.3	
Region - West (%)	13.8	37.4	48.9	
Carnegie-Research (%)	11.9	44.7	43.4	
Carnegie-Non Research (%)	8.3	38.3	53.3	
Carnegie-Uncl, PRO (%)	12.2	47.0	40.9	
Carnegie-Uncl, Other (%)	4.4	47.8	47.8	
Grant Award Characteristics				
NSF Award (%)	22.4	43.7	33.9	***
NIH Award	7.4	45.0	47.6	

Duration of Initial Award, Years	2.9	3.3	2.9

* p<.1; ** p<.05; ***p<.01 (p-values for chi square tests)

Table 3. Bivariate results: Data Sharing Status by Publication Counts

	Total	Archived	Informal	Not Shared
	n=935	n=111	n=415	n=409
	Median	Median	Median	Median
Primary Publications (w/ PI)	4	6	6	3
Primary Publications (w/ any Research Team Member)	5	8	6	3
Secondary Publications (no Team Member)	0	0	0	0
Total Publications	5	10	7	4
Total Publications including Students	2	4	3	1

Table 4. Multivariate Results: Negative Binomial Regression Models of Publication Counts

	Total # Publications, Self-Reported						Total # Secondary Publications, Self-Reported					
	Model 1			Model 2			Model 1			Model 2		
Data Sharing Status												
Primary Publications (w/ PI)	1.094	(0.123)	***	0.884	(0.128)	***	2.515	(0.415)	***	1.919	(0.443)	***
Shared Informally-Not Archived	1.020	(0.080)	***	0.837	(0.079)	***	2.375	(0.276)	***	1.565	(0.284)	***
Not Shared	ref			ref			ref			ref		
Primary Publications (w/any Research Team Member)												
PI Characteristics												
Age at award				0.025	(0.004)	***				0.037	(0.016)	**
Discipline - Health and Psychology				-0.254	(0.102)	**				-0.977	(0.370)	***
Discipline - Other (v.s Core Soc Sci)				-0.190	(0.130)					-1.107	(0.467)	***
Institutional Characteristics												
Carnegie-Non Res University				-0.685	(0.157)	***				-0.623	(0.584)	
Carnegie-Other				1.169	(0.246)	***				1.602	(0.840)	*
Carnegie-PRO (vs. Res Univ)				0.230	(0.113)					1.230	(0.387)	***
Grant Award Characteristics												
NIH (vs. NSF)				0.075	(0.093)					-0.202	(0.358)	
Duration of Award, Years				0.163	(0.027)	***				0.115	(0.102)	
Intercept	1.646	(0.058)		0.199	(0.222)		-1.314	(0.206)	***	-2.418	(0.794)	***
Dispersion	1.186			1.052			13.649			11.241		

log-likelihood estimates (standard errors in parentheses)

* p<.1; ** p<.05; ***p<.01

Table 5. Multivariate Results: Negative Binomial Regression Models of Publication Counts

	Total # Primary Publications, Self-Reported						Total # Student Publications, Self-Reported					
	Model 1			Model 2			Model 1			Model 2		
Data Sharing Status												
Primary Publications (w/ PI)	0.620	(0.111)	***	0.729	(0.112)	***	0.700	(0.156)	***	0.936	(0.165)	***
Shared Informally-Not Archived	0.743	(0.073)	***	0.67	(0.069)	***	0.763	(0.103)	***	0.665	(0.100)	***
Not Shared	ref			ref			ref			ref		
Primary Publications (w/any Research Team Member)												
PI Characteristics												
Age at award				0.024	(0.004)	***				0.022	(0.006)	**
Discipline - Health and Psychology				0.161	(0.089)	*				0.324	(0.125)	***
Discipline - Other (vs. Core Social Sci)				0.141	(0.113)					0.276	(0.165)	*
Institutional Characteristics												
Carnegie-Non Res University				-0.558	(0.142)	***				-0.888	(0.205)	***
Carnegie-Other				0.901	(0.211)	***				1.091	(0.335)	***
Carnegie-PRO (vs. Res Univ)				-0.216	(0.099)	**				-0.981	(0.147)	***
Grant Award Characteristics												
NIH (vs. NSF)				0.226	(0.081)	***				0.234	(0.113)	**
Duration of Award, Years				0.200	(0.024)	***				0.207	(0.034)	***
Intercept	1.444	(0.053)		-0.521	(0.206)		0.989	(0.075)	***	-0.982	(0.301)	***
Dispersion	0.902			0.75			1.832			1.559		

log-likelihood estimates (standard errors in parentheses)

* p<.1; ** p<.05; ***p<.01