NBER WORKING PAPER SERIES

MAN VERSUS MACHINE? SELF-REPORTS VERSUS ALGORITHMIC MEASUREMENT OF PUBLICATIONS

Xuan Jiang
Wan-Ying Chang
Bruce A. Weinberg

Working Paper 28431
http://www.nber.org/papers/w28431

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
February 2021

This work was supported by the National Science Foundation [contract numbers NSFDACS16T1400, NSFDACS16C1234] and by P01 AG039347. Weinberg was paid directly by NBER on P01 AG039347 and his work was supported on it through a subcontract from NBER to Ohio State. Any opinions, findings, and conclusions expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation or its staff. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Man Versus Machine? Self-Reports Versus Algorithmic Measurement of Publications
Xuan Jiang, Wan-Ying Chang, and Bruce A. Weinberg
NBER Working Paper No. 28431
February 2021
JEL No. C26,J24,J3,O31

ABSTRACT

This paper uses newly available data from Web of Science on publications matched to researchers in Survey of Doctorate Recipients to compare scientific publications collected by surveys and algorithmic approaches. We aim to illustrate the different types of measurement errors in self-reported and machine-generated data by estimating how publication measures from the two approaches are related to career outcomes (e.g. salaries, placements, and faculty rankings). We find that the potential biases in the self-reports are smaller relative to the algorithmic data. Moreover, the errors in the two approaches are quite intuitive: the measurement errors of the algorithmic data are mainly due to the accuracy of matching, which primarily depends on the frequency of names and the data that was available to make matches; while the noise in self reports is expected to increase over the career as researchers' publication records become more complex, harder to recall, and less immediately relevant for career progress. This paper provides methodological suggestion on evaluating the quality and advantages of two approaches to data construction. It also provides guidance on how to use the new linked data.

Xuan Jiang
The Ohio State University
1945 N High St
Columbus, OH 43210
jiang.445@osu.edu

Wan-Ying Chang
National Science Foundation
4201 Wilson Boulevard
Arlington, VA 22230
wchang@nsf.gov

Bruce A. Weinberg
The Ohio State University
Department of Economics
410 Arps Hall
1945 North High Street
Columbus, OH 43210
and NBER
weinberg.27@osu.edu

A data appendix is available at http://www.nber.org/data-appendix/w28431

1 Introduction

As academic maxim goes, “Publish or perish.” In this environment, it is important to be able to measure the publication trajectories of scientists; how they vary across researchers, including by gender, race, ethnicity, immigrant status, and over the career; and how they relate to outcomes such as earnings and placements. Traditionally, researchers have turned to survey questions on productivity, but recently researchers have focused on using algorithmic approaches. However, little is known about the validity and accuracy these two sources of data.

Among the best survey data available for such analyses have come from the Survey of Doctorate Recipients (SDR). The National Center for Science and Engineering Statistics (NCSES) within the National Science Foundation has conducted SDR biennially since 1973, collecting demographic information along with educational and occupational histories. Starting in 1995, self-reported data on publication counts for a reference period of five years are collected. On the other side, there are an increasing number of algorithmically disambiguated databases available commercially and from researchers. Examples include Elsevier, Microsoft Academic Graph, CiteSEER, Author-ity (Torvik et al., 2005; Torvik and Smalheiser, 2009). One of the best-known commercial databases is the Web of Science.

In recent years, NCSES has explored enriching survey data with alternative data sources to increase the value of its data and to avoid imposing further response burden of the survey participants. One aspect of this work was an effort to link survey respondents from the 1993-2013 SDR to publications indexed by the Web of Science (Freyman et. al, 2017). These linked data provide a unique opportunity to compare algorithmically disambiguated and matched data to survey data, including as determinants of research outcomes (Chang et al., 2019; Hopkins et al., 2013; Sabharwal, 2013; Whittington, 2011; Ginther et al., 2020).

We use these newly available data on publications matched to the SDR to compare survey-based and algorithm-based measures on the publications of scientists. Our work makes four contributions. First, we study how publication counts compare across the two approaches overall and for specific demographic groups and assess the relative quality of the two approaches. While our specific results only apply to the WoS data and some may prefer other publication and citation datasets, the WoS data are a prominent, respected dataset and our analysis provides a valuable head-to-head comparison of one algorithmic approach to the survey approach to collect complex data that may be informative. Second, our analysis involves estimating how publication measures from the two approaches are related to career outcomes (e.g. salaries, placements, and faculty rankings), estimates that are of interest in their own right. Third, as algorithmic approaches proliferate[1], so does the importance and cost of validation. While the conventional approach would be to collect a "gold standard," something that is costly and extremely difficult to do well (e.g. even CV data contain mistakes), we lay out a radically different and far lower cost approach to assessing data quality that others may find valuable. The utility of our approach has begun to be recognized, with Ginther et al. (2020) already applying it.[2] Last, we provide guidance on how to use the new linked SDR-WoS data.

It is important to realize at the outset that there are reasons to believe that both approaches to data collection are likely to yield different information, results, and generate some errors. First,

[1] The use of algorithmic data has recently growing in many research fields, including economics, social sciences, etc. For example, using geocoding and surname to impute race or using first name to impute gender is usual when demographic information is hard to be collected or matched (e.g. Marschke et al., 2018).

[2] As indicated, Ginther, Zambrana, and Chang (2020) start from our basic approach and modify it by adding a manually-collected gold standard. Their basic approach answers a slightly different question from ours. Their work is best suited for assessing the signal-to-noise ratio in the WoS, which is more useful for assessing the quality of the WoS and working with the WoS data. Our approach compares the signal-to-noise ratio in the WoS to that in the SDR self-reports. Thus, it is better suited to assisting a researcher who is choosing between using one dataset relative or the other. We also point to potential solutions. At a methodological level, their approach requires the development of a gold standard, which is difficult to do accurately and costly (as a consequence their main estimates are based on a sample that is roughly 7% the size of ours).

there is likely to be pure measurement error in retrospective self-reports. Second, self-reported publication counts, such as those included in the SDR typically do not account for the quality of publications and are, therefore, a noisy measure of scientific contributions. Lastly, it is possible that there is, what we might refer to as "bragging bias" in self-reports, with people who over-report their publications also over-reporting other outcomes such as salaries (bragging may well be less common for other outcomes, such as tenure status and institutional affiliation, which are more discrete). Another source of discrepancy between algorithmic data and self-reports is coverage, with the target coverage of articles and journals in a database being defined, while self-reports cover all journal publications. (Of course, this defined coverage may or may not provide more uniformly quality.)

The algorithmic approach has limitations and errors too, the most obvious of which are disambiguation and linking errors, which are likely to be greatest for relatively common names and when links must be made based on less information (e.g. without using e-mail addresses). Fortunately, for our comparison because disambiguation and linking errors and coverage issues stem from such different sources than self-reports, we have no reason to believe that the differences between the algorithmic data and self-reports will be highly correlated. Moreover, unlike the possible mis-reporting in the self-reports, the algorithmic measurement errors in the SDR-WoS dataset are unlikely to be correlated with over-reports in outcomes. Our methods section outlines a series of statistical approaches to assess the accuracy of the two data approaches.

Our findings are twofold. First, the overall results indicate that the potential biases in the SDR self-reports are likely smaller than those in the algorithmic data. While machine learning algorithms are improving rapidly over time and the WoS is only one algorithmic approach, this finding suggests that machine learning should not be viewed as categorically superior to self-

reports. Second, while we find that overall the algorithmic approach underperforms self-reports, the errors in the two approaches are quite intuitive, which could be leveraged by researchers using the SDR data and guide users and producers of other data. Specifically, the accuracy of matching depends on the frequency of names and the data that was available to make matches (e.g. e-mail addresses). By subsetting the sample, we are able to show that the algorithmic data perform as well or better than the self-reports for people with uncommon names or for whom more data were available for matching. On the other hand, the noise in self reports is expected to increase over the career as researchers' publication records become more complex, harder to recall, and less immediately relevant for career progress (e.g. once researchers have permanent positions and have progressed up the academic career ladder).

Taken together, our findings suggest potential approaches for users of the SDR publication data. One natural approach is to run estimates using both sets of variables and triangulate the findings. Depending on the research question, a second option would be to subset the data to the groups for which a given measure is more precise. A third option is to instrument for one measure (e.g. the self-reported publications) using the other (e.g. the algorithmic measure). Obviously, the preferred approach will depend on the specific research question.

The algorithmic approach has at least two additional advantages. First, the direct linkage to publications makes it possible to include a range of measures of scientific impact (i.e. citations to articles), which would be hard to collect as part of a survey. These can be used to control for publication quality, which we have explored, but including WoS citation / impact measures does not have a meaningful effect on our wage, affiliation, or promotion estimates. Second and as indicated, the algorithmic approach has the advantage of being lower burden for respondents, which might reduce the length of the survey or provide space for additional questions.

2 Data

2.1 SDR and Thomson Reuters Web of Science™

The Survey of Doctorate Recipients (SDR) is a longitudinal survey of individuals who have earned a research doctorate in science, engineering, or health (SEH) field from a U.S. institution. It is conducted by the National Center for Science and Engineering Statistics (NCSES) and contains a range of information on demographics, employment, and educational histories. The survey included questions about research outputs, including publications and patents, in 1995, 2003, and 2008.

To improve the understanding of the products of research by U.S.-trained SEH doctorates, NSF collaborated with Clarivate Analytics to use a machine learning approach to match the SDR respondents to the authors of publications indexed by the WoS. Drawing heavily on the description in Baker and Wolcott (2016), the matching algorithm incorporates name commonality, research field, education and employment affiliations, co-authorship network and self-citations to predict matches from the SDR respondents to the WoS. The overall procedure consists of five steps (Chang et. al, 2019):

1. A gold standard data set was constructed for use in training of prediction models and for validation of predicted matches;
2. Candidate publications were identified and blocked using last name and first initial;
3. The round one matching is conducted by Random ForestTM (RF) classification models trained to identify publications which could be matched to SDR respondents with a high degree of confidence. The high confident matches are called the ‘seed publications’ and used to increase the amount of data available for the subsequent matching;

4. Data were extracted from the seed publications and combined with survey data used in round one to enrich the RF models for increased recall and make the final predictions;

5. The final matched data set was refined to ensure that no respondent was matched to more than one authorship on a single publication and that those with an exact match by email were considered matches.

2.2 Subset of SDR and WoS Data Used in This Paper

For our purpose to assess and compare the validity and quality of the SDR self-reported and WoS-linked publications, we analyze all three waves of SDR that have been linked to WoS, 1995, 2003, and 2008 wave. We present the results from the 2003 survey, the middle of the three waves with publication data. Results for the 1995 and 2008 rounds are in the appendix. The results are qualitatively similar across all three waves of data, with exceptions noted below.

Our main dataset, 2003 SDR, provides individual-level, self-reports of journal article publication counts during 1998—2003[3]. We compare this publication to the aggregated number of publications from the WoS under the same time frame. The SDR also provides comprehensive information on doctorate recipients, including demographic characteristics, academic records, and career outcomes. We focus on the sample of doctorate recipients who placed in academia. Table 1 shows the summary statistics for the key variables we used in this paper: publication counts from SDR and WoS, self-reported salary, total annual federal research funding received by the employing institution, academic age, and gender share in the sample. The three outcome variables to measure individual career achievement are self-reported salary from SDR, total annual federal

[3] SDR surveys both journal article publication counts and conference article publication counts. We only use journal article publication counts in this analysis.

research funding received by the employment institution from NSF's Higher Education Research and Development (HERD) Survey[4], and tenure status[5]. The self-reported salary is in USD of the survey year (i.e., 2003, for the dataset we use in the main text). The lower panel of Table 1 shows the comparison of SDR and WoS publication counts by sub-category, such as gender, race, and faculty ranking. On average, WoS publication counts exceed SDR publication counts across the entire sample and all subgroups. The correlation coefficient indicates that they are somewhat but not highly correlated. As it is better to visualize the publication counts by academic age[6], we plot the average SDR and WoS publication by academic age, shown in Figure 1. It indicates that the WoS publication counts exceed those from the SDR at almost every academic age, except for the very senior ones.

3 Methods

The heart of our work will be to employ a standard set of econometric methods to address errors in the publication variables (Greene, 2003). This approach, which we believe is more broadly applicable differs markedly from a more conventional comparison to, for instance, a manually constructed "gold standard." Specifically, we analyze the relationship between faculty's self-reported career outcomes and the number of publications from both the SDR and WoS, to assess the quality of each measure of publication records. For our test case, the career outcomes, y_i, for

[4] The Higher Education Research and Development Survey is the primary source of information on R&D expenditures at U.S. colleges and universities. The survey collects information on R&D expenditures by field of research and source of funds and also gathers information on types of research, expenses, and headcounts of R&D personnel. The survey is an annual census of institutions that expended at least $150,000 in separately accounted for R&D in the fiscal year.

[5] Tenure status used in this study is defined by self-reported faculty rank in SDR. There are four faculty ranks: Assitant Professor, Associate Professor, Professor, and Other Faculty and Postdocs. We categorize Associate Professor and Professor as tenure = 1; and the rest as tenure = 0.

[6] Academic age is defined by number of years since the researcher got the Ph.D. degree.

individual i, include self-reported earnings, academic promotions, academic rank, and institution rank. Outcomes are related to researchers' own characteristics, X_i; including demographics (e.g. gender, ethnicity, immigrant status, and age), experience, and field of study. This section lays out our conceptual framework and approach.

We start from the assumption that the true number of publications by researcher i is P_i. Unfortunately, we do not observe the true number of publications. Rather, we observe the self-reported number of publications from SDR, P_i^{SDR}, and the number of matched publications from the WoS, $P_i{}^{WoS}$. We can write P_i^{SDR} and P_i^{WoS} as functions of P_i:

$$P_i^{SDR} = P_i + \varepsilon_i^{SDR} \quad (1A)$$

$$P_i^{WoS} = P_i + \varepsilon_i^{WoS} \quad (1B)$$

Where ε_i^{SDR} and ε_i^{WoS} are measurement errors in the SDR number of publication and WoS number of publications, respectively, which we regard as "errors-in-variables". These are assumed to be uncorrelated with true publications, $Cov\left(P_i, \varepsilon_i^{SDR}\right) = Cov\left(P_i, \varepsilon_i^{WoS}\right) = 0$. At the same time, it is possible that measurement error in the SDR self-reports might be correlated, most likely positively, with the true number of publications, $Cov\left(P_i, \varepsilon_i^{SDR}\right) \neq 0$. We discuss the sensitivity of our results to this violation of our assumptions. The main source of error in the SDR is reporting error on the part of respondents, while the main source of measurement error in the WoS data is due to algorithmic disambiguation and linking errors. Thus, it seems plausible these two types of measurement errors would be uncorrelated with each other, that $cov(\varepsilon_i^{SDR}, \varepsilon_i^{WoS}) = 0$.

For much of what we do, we condition on a range of control variables, in this case, we augment equations (1A) and (1B) by including control variables represented by X_i. In this formulation, we assume that $Cov\left(P_i, \varepsilon_i^{SDR}|X_i\right) = Cov\left(P_i, \varepsilon_i^{WoS}|X_i\right) = 0$; and that $cov(\varepsilon_i^{SDR}, \varepsilon_i^{WoS}|X_i) = 0$.

3.1 Horse Race Test: Compare Coefficients

A "horse race" test is a simple, informal approach to compare the ability (i.e. explanatory power) of the two key measures of publication counts to explain wages. Given that both measures are supposed to capture the same underlying construct, if one of the measures is a better predictor in the sense that it has a greater coefficient and stronger statistical significance, that indicates that it is measured more precisely (Farber and Gibbons, 1996, and Altonji and Pierret, 2001, employ similar logic, albeit in a very different context).

Specifically, consider a regression of some outcome y_i, such as the natural logarithm of wages, on publications. With the true data on publications, that regression would be

$$y_i = \beta P_i + \gamma X_i + u_i \quad (2A)$$

and it would say how an increased number of publications translates into higher wages. (Although it is not critical for our analysis, would be hesitant to give β a causal interpretation because wages and publications are both likely to be positively related to unobserved ability or motivation.)

Because we do not observe actual publications, we regress our outcomes, y_i, on the SDR and WoS measures of number of publications both one at a time (i.e. separate regressions) and also include them in the same regression. Our data are cross-sectional, and allow us to control for individual demographic and academic characteristics.

$$y_i = \beta^{SDR} P_i^{SDR} + \gamma X_i + u_i^{SDR} \qquad (2B)$$

$$y_i = \beta^{WoS} P_i^{WoS} + \gamma X_i + u_i^{WoS} \qquad (2C)$$

$$y_i = \beta^{SDR} P_i^{SDR} + \beta^{WoS} P_i^{WoS} + \gamma X_i + u_i^{Both} \qquad (2D)$$

and compare the coefficients, β^{SDR} and β^{WoS}. We note that the various β^{SDR} and β^{WoS} and γ estimates from (2B)-(2D) are analogous, but different parameters.

Under the assumptions that measurement errors in our publication measures are uncorrelated with each other $Cov(\varepsilon_i^{SDR}, \varepsilon_i^{WoS}|X_i) = 0$ and with the error in (2A) $cov(u_i, \varepsilon_i^{SDR}|X_i) = cov(u_i, \varepsilon_i^{WoS}|X_i) = 0$, standard errors-in-variables logic allows us to get a sense of the amount of measurement error in the two measures (in (1A) and (1B)) from the estimated coefficients β^{SDR} and β^{WoS}(Greene, 2003). To lay out the effects of these errors-in-variables formally, consider one of the regressions (2B) or (2C) where we regress outcomes on P_i^j, where $j \in \{SDR, WoS\}$.

$$\widehat{\beta^j} = \frac{Cov(y_i, P_i^j|X_i)}{Var(P_i^j|X_i)} = \frac{Cov(y_i, P_i + \varepsilon_i^j|X_i)}{Var(P_i|X_i) + Var(\varepsilon_i^j|X_i)} = \beta \frac{Var(P_i|X_i)}{Var(P_i|X_i) + Var(\varepsilon_i^j|X_i)} \quad (3)$$

Thus, the error in the variable P_i^j biases β^j toward zero relative to the true β in (2A) with the downward bias depending directly on the amount of noise in the publication measure P_i^j relative to the true variation in publications. We refer to this expression as the "errors-in-variables" formula. In the case where $Cov(P_i, \varepsilon_i^{SDR}) \neq 0$, $\widehat{\beta^{SDR}} = \frac{\beta Var(P_i|X_i) + Cov(P_i, \varepsilon_i^{SDR}|X_i)}{Var(P_i|X_i) + Var(\varepsilon_i^{SDR}|X_i) + 2Cov(P_i, \varepsilon_i^{SDR}|X_i)}$, so the bias will not solely reflect the signal-to-noise ratio in P_i^{SDR}.

Returning to our maintained assumptions that $Cov(P_i, \varepsilon_i^{SDR}|X_i) = Cov(P_i, \varepsilon_i^{WoS}|X_i) = 0$ and that the measurement errors in our publication measures are uncorrelated with each other $cov(\varepsilon_i^{SDR}, \varepsilon_i^{WoS}) = 0$ and with the error in (2A) $cov(u_i, \varepsilon_i^{SDR}|X_i) = cov(u_i, \varepsilon_i^{WoS}|X_i) = 0$, we can compare (qualitatively) the amount of noise in each publication measure from the relative size of β^{SDR} and β^{WOS}. Formally,

$$\frac{\widehat{\beta^{SDR}}}{\widehat{\beta^{WOS}}} = \frac{Var(P_i|X_i) + Var(\varepsilon_i^{WoS}|X_i)}{Var(P_i|X_i) + Var(\varepsilon_i^{SDR}|X_i)} \quad (4)$$

We can also assess the variance in publications. This can be done by noting that if $cov(\varepsilon_i^{SDR}, \varepsilon_i^{WoS}|X_i) = 0$, then

$$cov(P_i^{SDR}, P_i^{WoS}|X_i) = cov(P_i + \varepsilon_i^{SDR}, P_i + \varepsilon_i^{WoS}|X_i) = Var(P_i|X_i) \quad (5)$$

We conduct three broad analyses – a baseline analysis, an analysis of different age groups, and an analysis based on the match quality in WoS, for which we have a series of direct measures.

There are four critical assumptions underlying this analysis — that the measurement errors in our publication measures are uncorrelated with each other, $cov(\varepsilon_i^{SDR}, \varepsilon_i^{WoS}|X_i) = 0$, that each of the measurement errors is uncorrelated with the error in (2A) $cov(u_i, \varepsilon_i^{SDR}|X_i) = cov(u_i, \varepsilon_i^{WoS}|X_i) = 0$, and that $cov(P_i, \varepsilon_i^{SDR}|X_i) = cov(P_i, \varepsilon_i^{WoS}|X_i) = 0$. As indicated, we believe that $cov(\varepsilon_i^{SDR}, \varepsilon_i^{WoS}|X_i) = 0$ is reasonable. We believe that assuming $cov(u_i, \varepsilon_i^{WoS}|X_i) = 0$ is also reasonable – a counter example would be that people with more ambiguous names might over- or under-report their earnings. Given that Asian names (especially Chinese and Korean names) are the most ambiguous, this cannot be excluded as a possibility, but it also seems remote, especially because our baseline regressions control for race and ethnicity. If one of these assumptions were to be violated, the assumption $cov(u_i, \varepsilon_i^{SDR}|X_i) = 0$ seems like a likely candidate. We have referred to $cov(u_i, \varepsilon_i^{SDR}|X_i) > 0$ as "bragging bias", in which people who over-report their earnings may also over-report their publications. Such a violation would tend to bias $\widehat{\beta^{SDR}}$ upward, which would make it appear to be less noisy than the errors-in-variables formula (given in equation (3)) would suggest. Formally, in this case,

$$\widehat{\beta^{SDR}} = \frac{Cov(y_i, P_i^{SDR}|X_i)}{Var(P_i^j|X_i)} = \frac{Cov(y_i, P_i + \varepsilon_i^{SDR}|X_i)}{Var(P_i|X_i) + Var(\varepsilon_i^{SDR}|X_i)}$$

$$= \beta \frac{Cov(u_i, \varepsilon_i^{SDR}) + \beta Var(P_i|X_i)}{Var(P_i|X_i) + Var(\varepsilon_i^{SDR}|X_i)} \quad (3')$$

As above, if $cov(P_i, \varepsilon_i^{SDR}|X_i) \neq 0$, then equation (4) will not hold and $cov(P_i^{SDR}, P_i^{WoS}|X_i) = cov(P_i + \varepsilon_i^{SDR}, P_i + \varepsilon_i^{WoS}|X_i) = Var(P_i|X_i) + cov(P_i, \varepsilon_i^{SDR}|X_i)$, which would overstate the variance in P_i.

3.2 Instrumental Variable Analysis

Another approach is to employ instrumental variables. Instrumental variable analysis is a standard approach in many of the social sciences to address correlations between independent variables and the error term in a regression such as (2A)-(2D) (Goldberger, 1971; Bowden and Turkington, 1984; Morgan, 1990). There are two sources of bias in our estimates of β. One is the errors-in-variables in our P_i^j, for which instrumental variables is ideal when there are two measures of a predictor variable that are subject to measurement errors, but with uncorrelated errors (i.e. $cov(\varepsilon_i^{SDR}, \varepsilon_i^{WoS}|X_i) = 0$). It is also possible that $\widehat{\beta^{SDR}}$ is biased upward because of bragging bias, whereby people who overreport their earnings also overreport their publications $cov(u_i, \varepsilon_i^{SDR}|X_i) > 0$. So long as $cov(u_i, \varepsilon_i^{WoS}|X_i) = 0$, instrumenting for P_i^{SDR} with P_i^{WoS} can also address this concern. (At the same time, publications and wages both likely reflect "ability" and motivation, which are omitted from the models. Instrumental variables will not address this source of bias, because ability is likely to be correlated with earnings and both measures of publications.)

Formally, we regress P_i^{SDR} on P_i^{WoS}:

$$P_i^{SDR} = \delta^{WoS} P_i^{WoS} + \pi X_i + v_i^{SDR} \quad (6)$$

Where the coefficient,

$$\widehat{\delta^{WoS}} = \frac{Cov\left(P_i^{SDR}, P_i^{WoS} \middle| X_i\right)}{Var\left(P_i^{WoS} \middle| X_i\right)} = \frac{Cov\left(P_i + \varepsilon_i^{SDR}, P_i + \varepsilon_i^{WoS} \middle| X_i\right)}{Var(P_i|X_i) + Var\left(\varepsilon_i^{WoS} \middle| X_i\right)}$$

$$= \frac{Var(P_i|X_i)}{Var(P_i|X_i) + Var\left(\varepsilon_i^{WoS} \middle| X_i\right)} \quad (7)$$

under the assumption that $\varepsilon_i^{SDR} and\ \varepsilon_i^{WoS}$ are uncorrelated. Thus, we can directly estimate the variance in the measurement error in the WoS publication measure relative to the variance in publications (i.e. a signal to noise ratio). The more measurement error, the greater the attenuation bias and the closer $\widehat{\delta^{WoS}}$ will be to zero. In the case where $Cov\left(P_i, \varepsilon_i^{SDR} | X_i\right) \neq 0$, then $\widehat{\delta^{WoS}} = \frac{Var(P_i|X_i) + Cov\left(P_i, \varepsilon_i^{SDR} \middle| X_i\right)}{Var(P_i|X_i) + Var\left(\varepsilon'^{WoS}_i \middle| X_i\right)}$ then $\widehat{\delta^{WoS}}$ will overstate the signal to noise ratio in the SDR self-reports.

Of course, the same procedure can be run in reverse to obtain a measure of the signal to noise ratio in P_i^{SDR}, which generates a symmetric set of results. Specifically, if we estimate

$$P_i^{WoS} = \delta^{SDR} P_i^{SDR} + \pi X_i + v_i^{WoS} \quad (8)$$

We obtain

$$\widehat{\delta^{SDR}} = \frac{Cov\left(P_i^{WoS}, P_i^{SDR} \middle| X_i\right)}{Var\left(P_i^{SDR} \middle| X_i\right)} = \frac{Cov\left(P_i + \varepsilon_i^{W0S}, P_i + \varepsilon_i^{SDR} \middle| X_i\right)}{Var(P_i|X_i) + Var\left(\varepsilon_i^{SDR} \middle| X_i\right)}$$

$$= \frac{Var(P_i|X_i)}{Var(P_i|X_i) + Var\left(\varepsilon_i^{SDR} \middle| X_i\right)} \quad (9)$$

This gives us an estimate of the measurement error in the SDR publication measure. In the case where $Cov\left(P_i, \varepsilon_i^{SDR} | X_i\right) \neq 0$, then $\widehat{\delta^{SDR}} = \frac{Var(P_i|X_i) + Cov\left(P_i, \varepsilon_i^{SDR} \middle| X_i\right)}{Var(P_i|X_i) + Var\left(\varepsilon_i^{WoS} \middle| X_i\right) + 2Cov\left(P_i, \varepsilon_i^{SDR} | X_i\right)}$ then $\widehat{\delta^{WoS}}$ will overstate the signal to noise ratio in the SDR self-reports.

Equation (6) and (8) above form the first stage of the instrumental variable estimation. The second stage of the instrumental variables procedure involves taking the prediction from the first stage and including it as an independent variable in a regression where the original outcome is the dependent variable. Specifically, we estimate models where outcomes are related to publications using both publication measures one by one as in equation (1) and using instrumental variables to address measurement error. Formally, we first estimate equation (6). We then regress the outcome variable, y_i, on the fitted value from the first stage, $\widehat{P_i^{SDR}}$

$$y_i = \beta^{SDR}\widehat{P_i^{SDR}} + \gamma X_i + u_i \qquad \text{(second stage)}$$

The estimates from this model will account for measurement error in P_i^{SDR} even if the measurement in P_i^{SDR} is correlated with the error in the earnings equation, with $\beta^{SDR} = \beta$ so long as the error in the earnings equation is not also correlated with the error in P_i^{WoS}, ε_i^{WoS}. Intuitively, the instrumental variables procedure uses the portion of P_i^{SDR} that is predicted by P_i^{WoS}, which is assumed to be uncorrelated with the error in the earnings equation.

We can also implement this procedure in reverse running equation (8) as our first stage and then regressing the outcome on the predicted value from that equation in a second stage. This model will produce unbiased estimates under assumptions that the measurement errors in the two publication measures are uncorrelated with each other and that the measurement error in P_i^{SDR}, ε_i^{SDR}, is uncorrelated with the measurement error in the outcome equation, u_i. Insofar as we may have some questions about the last assumption, this approach may be less compelling than the former. At the same time, under the assumptions that $cov(\varepsilon_i^{SDR}, \varepsilon_i^{WoS}|X_i) = 0$ and that $cov\left(u_i, \varepsilon_i^{SDR}|X_i\right) = cov(u_i, \varepsilon_i^{WoS}|X_i) = 0$, $\widehat{\beta^{SDR}} = \widehat{\beta^{WoS}} = \beta$. That is, if both measurement

errors are orthogonal to the error in the outcome equation, they will both be valid instruments and generate consistent estimates for β.

We note that if $Cov\left(P_i, \varepsilon_i^{SDR} | X_i\right) \neq 0$ then

$$\widehat{\beta^{SDR}} = \frac{Cov\left(Y_i, P_i^{WoS} \middle| X_i\right)}{Cov\left(P_i^{SDR}, P_i^{WoS} \middle| X_i\right)} = \frac{Cov(y_i, P_i + \varepsilon_i^{WOS} | X_i)}{Cov\left(P_i + \varepsilon_i^{SDR}, P_i + \varepsilon_i^{WoS} \middle| X_i\right)}$$

$$= \beta \frac{Cov(P_i, P_i + \varepsilon_i^{WOS} | X_i)}{Var(P_i | X_i) + Cov\left(P_i, \varepsilon_i^{SDR} | X_i\right)} = \beta \frac{Var(\, P_i | X_i)}{Var(P_i | X_i) + Cov\left(P_i, \varepsilon_i^{SDR} | X_i\right)}$$

Thus, if $Cov\left(P_i, \varepsilon_i^{SDR} | X_i\right) \neq 0$, $\widehat{\beta^{SDR}}$ will be biased downward with the bias being greater the greater the covariance between P_i and ε_i^{SDR}. Similarly, if $Cov\left(P_i, \varepsilon_i^{SDR} | X_i\right) \neq 0$ then

$$\widehat{\beta^{WoS}} = \frac{Cov\left(Y_i, P_i^{SDR} \middle| X_i\right)}{Cov\left(P_i^{WoS}, P_i^{SDR} \middle| X_i\right)} = \frac{Cov\left(y_i, P_i + \varepsilon_i^{SDR} \middle| X_i\right)}{Cov\left(P_i + \varepsilon_i^{WoS}, P_i + \varepsilon'^{SDR}_i \middle| X_i\right)}$$

$$= \beta \frac{Cov\left(P_i, P_i + \varepsilon_i^{SDR} \middle| X_i\right)}{Var(P_i | X_i) + Cov\left(P_i, \varepsilon_i^{SDR} | X_i\right)} = \beta \frac{Var(\, P_i | X_i) + Cov\left(P_i, \varepsilon_i^{SDR} | X_i\right)}{Var(P_i | X_i) + Cov\left(P_i, \varepsilon_i^{SDR} | X_i\right)} = \beta$$

Thus, $\widehat{\beta^{WoS}}$ is consistent even if $Cov\left(P_i, \varepsilon_i^{SDR} | X_i\right) \neq 0$. This result obtains because we the measurement error in the instrument biases the first stage and reduced form estimates similarly leaving the second stage estimates consistent.

4 Findings

4.1 Results of the Horse Race Tests

4.1.1 Baseline

We first assess the explanatory power of the two publication measures to our three outcome variables on the base of the entire sample. In specific, we estimate equation (1) using the full sample of those worked in academic sector in the 2003 SDR. Note that we replicate all analyzers using the 2008 and 1995 SDR as a robustness check. We estimate this model with three

functional forms for the independent variables of interest: 1. The SDR publication counts and the WoS publication counts, 2. ln(SDR count + 1) and ln(WoS count +1); and 3. ln(SDR count), ln(WoS count), and indicators of zero publications. In specific, when we use logarithmic measures of publications, i.e. ln(SDR count) and ln(WoS count), we set the log count variables equal to zero (ln(SDR count) = 0) if SDR count = 0 and control for indicators of zero publication for SDR. In this procedure, the indicator on having zero publications gives the difference in wages between people with zero and one publication.

Table 2 shows the main results of the baseline analysis with ln(salary) as the outcome variable. We explore several regression specifications (each column is a separate regression model). Model (1) only includes the key variable, SDR publication counts and WoS publication counts. Model (2) adds in demographic controls including, a female indicator, an indicator of marital status, a linear term of academic age, a square term of academic age, and race indicators. Model (3) adds in field of study fixed effects by a coarse definition of fields of study (7 fields in total). Model (4) replaces the coarse field of study fixed effects with a fine definition (83 fields in total). Model (5) replaces the linear and squared terms in academic age with a full set of academic age fixed effects, to further control more flexibly for variation in academic career stage. Model (5) is the richest model and is our preferred specification. Model (6) has the same specification as model (5) but replaces the SDR and WoS publication count measured in levels with ln(SDR+1) and ln(WoS +1). Model (7) replace them with ln(SDR), ln(WoS), and an indicator for zero SDR publications. The standard errors are clustered at academic age level.

In general, we find that the 1998-2003 SDR's publication count has substantially greater explanatory power for the 2003 base annual salary than WoS's publication count over the same time period. This relationship holds across all regression specifications: with or without

demographic characteristics, controlling for linear or nonlinear academic age, controlling for coarse or fine research field classifications, and estimated in levels or either of the natural logarithm specifications of publication count. Specifically, one SDR publication is associated with a 0.7% increase in salary while one WoS publication is associated with a 0.34% increase in salary. The explanatory power of the model that uses SDR publications is more than double that of the one that uses WoS publications.

We replicate this table using the 1995 and 2008 rounds of the SDR, which also include self-reported publication count, and find that the results are robust. Interestingly, the magnitudes of the coefficients on the SDR publications are larger in the 2008 and 2003 rounds than in the 1995 round. By contrast the magnitude of the coefficient on the WoS publications is quite similar across all three years. One potential explanation is that people are becoming better at keeping publication records and reducing the measurement errors in self-reported publication count. Another explanation is that the rate of return to publishing has increased, but that measurement error in the WoS has also increased, which we find less plausible because the quality of algorithmic disambiguation and linkage typically improve over time as more data elements are available.Although salaries are a very common measure of career outcome, there is concern that the measurement error in the self-reported salary and the measurement error in the independent variable—the SDR publication count—might be correlated. We then investigate alternative measures of career outcomes, which are presumably less subject to individual reporting errors. One of the alternative outcome variables is federal research funding received by respondents' institutions. This measure is an institution-level outcome, which is strongly related to reputational rankings but available for considerably more institutions. We use it as a proxy for

the quality of academic placement[7]. We estimate the same set of regressions in Table 2 replacing the dependent variable with the natural logarithm of federal research funding received by the employing institution. Estimates in Table 3 shows the same pattern as we see in Table 2. The coefficients of SDR counts are 2-3 times larger than the coefficients of WoS counts, although the differences between the two coefficients diminish when we use a log specification for publication count.

The third outcome variable is researcher's tenure status, which is defined by the self-reported faculty rank (i.e., assistant professor, associate professor, professor, or other faculty and postdocs). Faculty rank is self-reported and may be misreported, but because it is discrete and highly salient, it may be reported more accurately than salary, which is continuous and is known to be reported with error. Table 4 shows the estimates using a dummy variable of tenure status[8] as the dependent variable. The finding that SDR publication counts have more explanatory power than WoS counts is robust and even stronger than results in Table 2 and Table 3.

[7] Respondents were asked their institutional affiliations and we assigned funding amounts from NSF's HERD data, so measurement error would have to come from respondents misreporting their institutions or survey staff misrecording it, which seems unlikely to be widespread.

[8] The dummy variable tenure status is defined as 1 if faculty rank is professor or associate professor; 0 otherwise.

4.1.2 Academic Age Group Heterogeneity

We hypotheses that self-reported publication counts will become increasingly noisy over the career as a researcher has more publications and the precise timing of publications becomes harder to recall. If so, self-reported publication counts are likely to degrade in quality later in the career. By contrast, the accuracy of WoS's algorithmically generated publication counts should not change over the career. These hypotheses lead us to expect the coefficient on the SDR's self-reported publication counts to decline over the career while the coefficient on WoS's publication counts to be roughly stable, or perhaps increase in specifications where the coefficients on SDR publication counts decline. We run the horse race test by age quintiles. As shown in Table 5, we find that the coefficient on the SDR's publication count becomes less important as academic age increases, while the coefficient on the WoS's publication count becomes more important as academic age increases. In particular, the coefficients on the WoS's publication count is not statistically different from zero in quintile 1, where the dependent variable is ln(Salary), but significantly different from zero in quintile 2—5. Similarly, the coefficients on the WoS's publication count is not statistically different from zero in quintile 1—3, where the dependent variable is ln(Funding), but significantly different from zero in quintile 4—5.

4.1.3 Match Quality Analysis

The quality of self-reported publication counts are likely to vary according to a series of WoS's match quality measures. Specifically, if WoS was unable to make a high-quality match, the WoS publication counts are likely to be noisy and then the coefficient on the WoS publication counts should be lower. By contrast, we do not expect the accuracy of the SDR's self-reported publication counts to vary with WoS match quality and therefore we expect the

coefficients on the SDR's self-reports to be essentially constant or perhaps increase slightly as WoS match quality deteriorates and hence the coefficient on WOS publication counts declines.

We explored three publication-level measures on the match quality of WoS. First, the probability that a publication matches to the SDR respondent, as predicted by WoS's random forest models, which is a continuous variable ranging between (0.5, 1). The higher the match probability, the more accurate the match. Second, the round at which the match was made, which indicates the number of rounds the random forest took to match the publication to the respondent. This is a discrete variable taking values of 0, 1 and 2, representing descending match quality: matched with an email address (value = 0), matched on round 1 (value = 1), and matched on round 2 (value = 2). The lower the number of match rounds, the higher of the match quality. Both the match probability and match rounds are at the article level. To obtain an author level measure, we take the mean of each variable across each respondent's publications and use the author mean to proxy for WoS's match quality. Third, the frequency that a particular respondent's last name and first initial combination appears in the WoS database, which ranges from 1 to 218422. We expect that the lower the frequency, the higher the match quality. We estimate the horse race model across subsamples defined by quintile or quartile of each match quality proxy.

Table 6 shows different coefficients on SDR's publication counts and WoS's publication counts across quintiles of match probability. Each panel, from the top to the bottom, presents the regression estimates for each outcome variable: salary, funding, and tenure status. We find that: when the match probability is low (quintiles 1 or 2), WoS publication counts have significantly lower explanatory power than SDR publication counts. As the match quality gets higher, the coefficient on WoS publication counts increases and even exceeds the coefficient on SDR

publication counts in some subgroup. This pattern is robust in the replicated results using the 2008 and 1995 datasets.

A similar pattern emerges in Table 7, where the coefficient on the WoS publication counts is higher for people with more WoS publications that were matched in earlier rounds. Specifically, the coefficient on the WoS publication counts is not statistically different or even larger than the coefficient on SDR publication counts in the 1st and 2nd quintiles. This finding reaffirms that WoS publication counts has larger explanatory power than that of SDR when the WoS publication is matched with the highest quality.

Lastly, Table 8 shows that when the frequency of names identified for the same author is low (quintiles 1—3), the coefficient on WoS publication counts is mostly higher than the coefficient on SDR publication counts[9]. For authors of the publications with common / ambiguous names (i.e., quintiles 4—5), the SDR's coefficient is significantly larger than WoS's.

Thus, across our three analyses, we find considerable and plausibly differential explanatory power for WoS publication counts across different levels of match quality in the WoS data. In general, the quality of the algorithmic data appears to be as good as the self-reported data when we restrict to the highest quality matches, although the overall quality of the algorithmic data is not as good as the self-reports. Still with improvements in matching algorithms, it seems plausible that algorithmic approaches have or will surpass many self-reports.

4.2 Results of the Instrument Variable Analysis

We now compare the measurement errors in the two data sources using instrumental variables. We present the second stage and the first stage in the upper and lower panel of Table 9,

[9] We have no explaination of the coefficient of WoS in column (1) of last panel is negative.

respectively. We find both the deltas (coefficients in the first stage), δ^{SDR} and δ^{WoS}, and the betas (coefficients in the second stage), β^{SDR} and β^{WoS}, are statistically different. Specifically, β^{SDR} is statistically larger than β^{WoS}, which reinforces our previous finding in the horse race tests. Compared to the coefficients in OLS models, the coefficients, β^{SDR} and β^{WoS}, in the IV models are significantly larger (more than double) than the coefficients in the OLS analysis. This finding is important because it indicates that there is considerable measurement error in both P^{SDR} and P^{WoS}.

The magnitudes of the coefficients, δ^{SDR} *and* δ^{WoS}, from the first stage are also informative. We find that δ^{WoS} is near 0.4 across all specifications, which is consistent with substantial measurement error. By contrast, δ^{SDR} is close to 1. Specifically, the low δ coefficient in the first stage (i.e. $\hat{\delta} < 1$) shown in the lower panel of Table 9, is consistent with measurement error in publication counts. The lower estimate for δ^{WoS} compared to δ^{SDR} corroborates our results above, indicating that there is more measurement error in the WOS measures of publication counts than in the SDR measure.

One argument for using algorithmic data is that it is possible to obtain far greater information on the quality of articles from algorithmic links than from self-reports. In other analyses, we have taken advantage of the measures of article quality (citation counts and the quality of journals) that are available in the WoS. Specifically, we add one of the following variables as a measure of article quality: the total number of citations in the past five years, the average number of citations per year, the total impact of journals in the past five years, or the average impact of journals. These estimates do not show marked improvements in the explanatory power of the WoS

data, which suggests that even the additional wealth of information in WoS does not offset its greater noise.

5. Conclusions

Machine learning provides a promising route to data linkage, which efficiently increases the utility of individual data sources and offers a wealth of research opportunities. This paper has explored a case of importance to the scientific community in which both algorithmic and survey responses are available and can be compared. Specifically, we focus on measures of scientists' publications and how they relate to career outcomes measured by earnings, placements, and faculty rank. Perhaps surprisingly, we find that overall the publication counts generated by machine learning are more noisy than self-reports. Moreover, we find that the relative noise in the two sets of measures varies in an intuitive way. That is, self-reports degrade for senior researchers who have more publications and for whom it may be harder to recall the exact quantity and timing of publications. Algorithmic measures degrade when names are more ambiguous (i.e. more common) or less data is available to make high-quality matches.

These findings are valuable for researchers using the publication variables in the SDR. A first approach is to run estimates using both sets of variables and triangulates the findings. Second, depending on the research question, one might subset the data to the groups for which a given measure is more precise. A third option is to instrument for one measure (e.g. the self-reported measure) using the other (e.g. the algorithmic measure). Obviously, the preferred approach will depend on the specific research question.

Other known databases, such as Google Scholar, Microsoft Academic Graph, or Scopus, might have different coverage of publications and/or authors and/or take a different

disambiguation approach, and thus have different quality (Paszcza, 2016). Our analysis is based on the WoS database, which is the only database for which links to the SDR self-reported publications are available. Thus, while the use of one specific dataset is a limitation, the WoS is a prominent and widely-used dataset and we believe that the analysis contributes to the scarce literature on comparing self-reported and algorithmic data.

Stepping back from our specific context, our results may be useful for data producers. They suggest that despite the understandable enthusiasm for them, even high-quality algorithmic approaches are not yet uniformly superior to self-reports. At the same time, we expect algorithmic approaches to improve over time relative to survey methods. Another promising approach that might be explored are recommendation systems, where algorithmic methods are used to populate lists for people to accept or reject. The ORCiD system (https://orcid.org/) is one prominent example of such an approach. Such approaches also have the potential to reduce burdens on survey respondents at the same time that they allow for training data that can be used to further refine algorithmic approaches.

We also believe that our underlying approach is broadly applicable. The conventional approach to assessing data quality is to manually build a "gold standard," which can be quite time-consuming and costly. Our approach permits estimation of data quality without requiring a gold standard using accepted statistical methods.

Acknowledgement

This work was supported by the National Science Foundation [contract numbers NSFDACS16T1400, NSFDACS16C1234] and by P01 AG039347. Weinberg was paid directly by NBER on P01 AG039347 and his work was supported on it through a subcontract from NBER to Ohio State. Any opinions, findings, and conclusions expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation or its staff.

References

Altonji, JG, Pierret, CR. Employer Learning and Statistical Discrimination. Quarterly Journal of Economics. 2001 Feb;116(1):313-50.

Baker C and Wolcott H, NSF SDR Matching to Publications & Patents, Final Report, Research Data Science & Evaluation, Thomson Reuters, August 4, 2016

Breiman, Leo (2001), "Random forests." 5–32.

Bowden, R. J., & Turkington, D. A. (1990). Instrumental variables (Vol. 8). Cambridge university press.

Chang WY, White KE, and Sugimoto CR, Demographic Differences in the Publication Output of U.S. Doctorate Recipients, in 7th Intern. Conf. on Scientometrics & Informetrics (ISSI), Rome Italy, Sept. 2019, pp. 2430-2439

Farber, HS, Gibbons R. Learning and Wage Dynamics. Quarterly Journal of Economics. 1996 Nov; 111(4):1007-47.

Freyman C, Deitz S, Ross L, Benskin J, Kalathil N, and Davis T, Manual Evaluation of Scientific Productivity Data of SDR Sample through Linkage, NCSES Task Order 09, Results from the manual matching evaluation. 2017

Gerald Marschke, Allison Nunez, Bruce A. Weinberg, and Huifeng Yu. 2018. "Last Place? The Intersection of Ethnicity, Gender, and Race in Biomedical Authorship" *American Economic Review Papers and Proceedings* 108 (No. 5, May): 222-27.

Goldberger, A. S. (1972). Structural equation methods in the social sciences. Econometrica: Journal of the Econometric Society, 979-1001.

Greene WH, Econometric analysis. Pearson Education India. 2003.

Hopkins, Allison, James Jawitz, Christopher McCarty, Alex Goldman, and Nandita Basu (2013), "Disparities in publication patterns by gender, race and ethnicity based on a survey of a random sample of authors." Scientometrics, 96 (2), 515–534.

Morgan, M. S. (1990). The history of econometric ideas. Cambridge University Press.

Paszcza, Bartosz. Comparison of Microsoft academic (graph) with web of science, scopus and google scholar. Diss. University of Southampton, 2016.

Sabharwal, Meghna (2013), "Comparing research productivity across disciplines and career stages." Journal of Comparative Policy Analysis: Research and Practice, 15, 141–163.

Torvik VI, Smalheiser NR. Author name disambiguation in MEDLINE. ACM Transactions on Knowledge Discovery from Data (TKDD). 2009 Jul 28;3(3):1-29.

Torvik VI, Weeber M, Swanson DR, Smalheiser NR. A probabilistic similarity metric for Medline records: A model for author name disambiguation. Journal of the American Society for information science and technology. 2005 Jan 15;56(2):140-58.

Whittington, Kjersten (2011), "Mothers of invention?: Gender, motherhood, and new dimensions of productivity in the science profession." Work and Occupations, 38, 417–456.

Tables and Figures

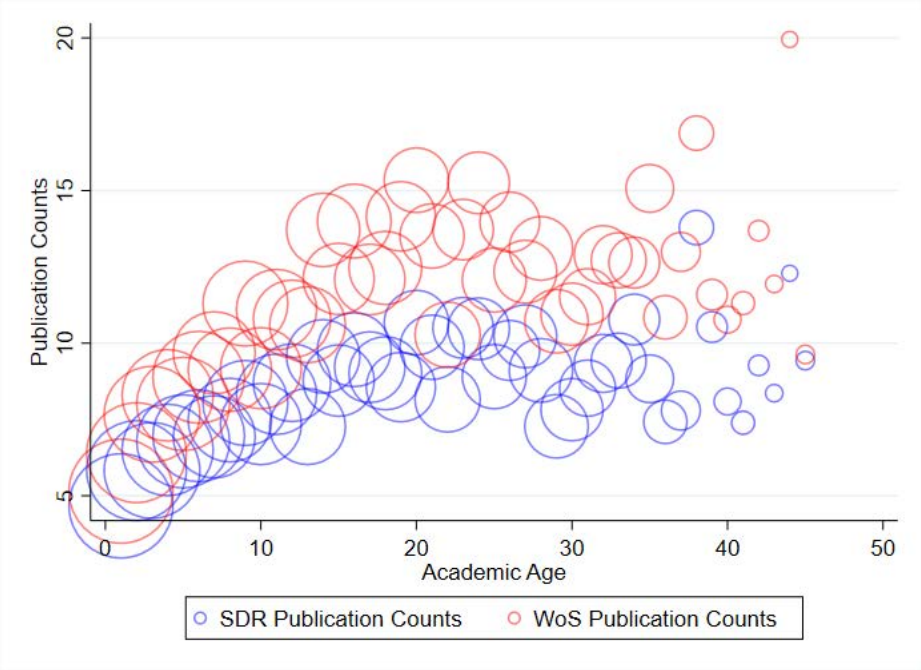

Figure 1: SDR v.s. WoS Publication Counts across Academic Age

Notes: This figure shows the average publication counts by academic age from the two sources: SDR and WoS in the period of 1998--2003. The size of each bubble represents the size of the academic age group. Survey weights are used in creating this figure.

Table 1. Summary Statistics

Variable	Obs	Weight	Mean	Std. Dev.
SDR_Pub	10,738	238053.4	8.049375	11.26069
WoS_Pub	10,738	238053.4	10.78217	18.11434
Salary	10,738	238053.4	74748.89	44173.57
Federal Research Funding by Institution	6,435	141374.8	209320.5	216959.2
Academic Age	10,716	237487.3	20.47783	10.84935
Correlation Coefficient between SDR&WoS	0.6838			
Male				
SDR_Pub	3,758	70938.51	6.263584	8.358401
WoS_Pub	3,758	70938.51	8.052527	14.97344
Female				
SDR_Pub	6,980	167114.9	8.807425	12.20892
WoS_Pub	6,980	167114.9	11.94088	19.17743
Race = Asian				
SDR_Pub	739	7487.479	7.88612	11.09726
WoS_Pub	739	7487.479	8.477628	13.82493
Race = Black				
SDR_Pub	637	7262.115	4.480074	5.7293
WoS_Pub	637	7262.115	5.241426	10.04866
Race = Hispanic				
SDR_Pub	1,344	29532.64	9.272687	11.70309
WoS_Pub	1,344	29532.64	10.759	18.2453
Race = Others				
SDR_Pub	7,799	190300.7	7.981884	11.26394
WoS_Pub	7,799	190300.7	11.08077	18.37407
Race = White				
SDR_Pub	219	3470.463	9.161368	14.22251
WoS_Pub	219	3470.463	11.17215	21.52893
Facultyrank_ordered = "Other Faculty and Postdoc"				
SDR_Pub	2,379	51134.9	4.970001	7.745157
WoS_Pub	2,379	51134.9	7.063239	12.43537
Facultyrank_ordered = "Assistant Professor"				
SDR_Pub	2,352	46999.94	6.828229	7.450015
WoS_Pub	2,352	46999.94	8.813287	13.15191
Facultyrank_ordered = "Associate Professor"				
SDR_Pub	2,316	50982.98	7.533814	9.269903
WoS_Pub	2,316	50982.98	9.727193	13.82979
Facultyrank_ordered = "Professor"				
SDR_Pub	3,341	81509.44	11.48489	14.96245
WoS_Pub	3,341	81509.44	15.48039	24.45626

Notes: This table shows summary statistics of the key variables used in this paper from the 2003 SDR and respondent's SDR and WoS publication counts between 1998—2003.

Table 2. Horse Race Test: Salary

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
SDR_Pub	0.00952***	0.00796***	0.00807***	0.00842***	0.00839***		
	(0.0008)	(0.0007)	(0.0008)	(0.0008)	(0.0008)		
WoS_Pub	0.00437***	0.00334***	0.00350***	0.00327***	0.00328***		
	(0.0006)	(0.0004)	(0.00050)	(0.0005)	(0.0005)		
ln_SDR_Pub1						0.10962***	
						(0.0120)	
ln_WoS_Pub1						0.05231***	
						(0.0074)	
ln_SDR_Pub							0.08513***
							(0.0095)
ln_WoS_Pub							0.05530***
							(0.0070)
Zero_SDR_Pub=1							-0.05264*
							(0.0264)
Female		-0.08485***	-0.08146***	-0.07375***	-0.07258***	-0.05943***	-0.05955***
		(0.0143)	(0.0160)	(0.0169)	-0.0169)	-0.0164)	(0.0165)
Marital status		0.0048	0.0045	0.0047	0.0122	0.0131	0.0126
		(0.0134)	(0.0135)	(0.0135)	(0.0155)	(0.0155)	(0.0154)
Academic Age		0.04248***	0.04242***	0.04186***			
		(0.0032)	(0.0032	(0.0032			
Academic Age2		-0.00071***	-0.00069***	-0.00067***			
		(0.0001)	(0.0001)	(0.0001)			
Asian		0.0210	0.0188	0.0217	0.0193	0.0160	0.0161
		(0.0451)	(0.0441)	(0.0439)	(0.0428)	(0.0437)	(0.0442)
Black		0.0844	0.0687	0.0639	0.0604	0.0775	0.0764
		(0.0538)	(0.0548)	(0.0562)	(0.0567)	(0.0556)	(0.0562)
Hispanic		0.0527	0.0457	0.0391	0.0395	0.0336	0.0331
		(0.0445)	(0.0437)	(0.0418)	(0.0423)	(0.0424)	(0.0426)
Others		0.0367	0.0365	0.0391	0.0388	0.0320	0.0329
		(0.0422)	(0.0413)	(0.0414)	(0.0415)	(0.0418)	(0.0424)
Field of Study FE (Coarse)	NO	NO	YES	NO	NO	NO	NO
Field of Study FE (Fine)	NO	NO	NO	YES	YES	YES	YES
Academic Age FE	NO	NO	NO	NO	YES	YES	YES
Observations	10716	10101	10101	10101	10101	10101	10101
R-squared	0.0493	0.1315	0.1420	0.1581	0.1642	0.1761	0.1759
Adjusted R-squared	0.0492	0.1306	0.1407	0.1502	0.1528	0.1648	0.1644

Notes: The dependent variable in this table is the natural log of the self-reported salary in 2003. Each column is a separate regression, with different set of control variables. Column (1) include level of SDR publication count and level of WoS publication count. Column (2) add in demographic controls including, an indicator of female, an indicator of marital status, a linear trend of academic age, a square term of academic age, and race indicators. Column (3) add in field of study fixed effects by a coarse definition of fields of study (7 fields). Column (4) replace the field of study fixed effects from the one by the coarse definition to a fine definition (83 fields). Column (5) replace the linear and squared terms of academic age to a full set of academic age fixed effects. Column (5) is our preferred specification. Column (6) has the same specification as column (5) but replace level SDR and WoS publication counts with log(SDR+1) and log(WoS +1). Column (7) replace them with log(SDR), log(WoS), and an indicator for zero SDR. The standard errors are clustered at academic age level. *p<0.1, **p<0.05, ***p<0.01.

Table 3. Horse Race Test: Federal Research Funding

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
SDR_Pub	0.02971***	0.03049***	0.03089***	0.03130***	0.03124***		
	(0.0030)	(0.0031)	(0.0029)	(0.0029)	(0.0029)		
WoS_Pub	0.01481***	0.01555***	0.01314***	0.01164***	0.01146***		
	(0.0020)	(0.0019)	(0.0018)	(0.0017)	(0.0017)		
ln_SDR_Pub1						0.34913***	
						(0.0311)	
ln_WoS_Pub1						0.33929***	
						(0.0386)	
ln_SDR_Pub							0.31932***
							(0.0407)
ln_WoS_Pub							0.31453***
							(0.0373)
Zero_SDR_Pub=1							0.0383
							(0.1288)
Female		0.20885***	0.21334***	0.15406**	0.16265**	0.21717***	0.21848***
		(0.0579)	(0.0624)	(0.0603)	(0.0612)	(0.0636)	(0.0621)
Marital status		0.0143	0.0193	0.0283	-0.0333	-0.0339	-0.0359
		(0.0470)	(0.0461)	(0.0455)	(0.0512)	(0.0486)	(0.0488)
Academic Age		-0.06806***	-0.06931***	-0.07085***			
		(0.0133)	(0.0127)	(0.0123)			
Academic Age2		0.00164***	0.00172***	0.00182***			
		(0.0003)	(0.0003)	(0.0003)			
Asian		-0.0949	-0.1411	0.0562	0.0445	-0.0307	-0.0262
		(0.2218)	(0.2299)	(0.2257)	(0.2176)	(0.2024)	(0.1990)
Black		0.39291*	0.3753	(0.3263	0.3358	-0.1934	-0.1942
		(0.2239)	(0.2425)	(0.2436)	(0.2420)	(0.2246)	(0.2233)
Hispanic		0.53818**	0.44568*	0.43972*	0.42763*	0.42464**	0.42064**
		(0.2121)	(0.2254)	(0.2232)	(0.2163)	(0.2056)	(0.2035)
Others		0.0343	0.0021	0.0849	0.0773	0.0360	0.0416
		(0.1980)	(0.2079)	(0.2061)	(0.1982)	(0.1825)	(0.1815)
Field of Study FE (Coarse)	NO	NO	YES	NO	NO	NO	NO
Field of Study FE (Fine)	NO	NO	NO	YES	YES	YES	YES
Academic Age FE	NO	NO	NO	NO	YES	YES	YES
Observations	6420	6059	6059	6059	6059	6059	6059
R-squared	0.0587	0.0784	0.1051	0.1445	0.1567	0.1940	0.1957
Adjusted R-squared	0.0584	0.0768	0.1028	0.1312	0.1374	0.1757	0.1769

Notes: The dependent variable in this table is the natural log of federal research funding received by the employing institution in 2003. Each column is a separate regression, with different set of control variables. Column (1) include level of SDR publication count and level of WoS publication count. Column (2) add in demographic controls including, an indicator of female, an indicator of marital status, a linear trend of academic age, a square term of academic age, and race indicators. Column (3) add in field of study fixed effects by a coarse definition of fields of study (7 fields). Column (4) replace the field of study fixed effects from the one by the coarse definition to a fine definition (83 fields). Column (5) replace the linear and squared terms of academic age to a full set of academic age fixed effects. Column (5) is our preferred specification. Column (6) has the same specification as column (5) but replace level SDR and WoS publication counts with log(SDR+1) and log(WoS +1). Column (7) replace them with log(SDR), log(WoS), and an indicator for zero SDR. The standard errors are clustered at academic age level. *p<0.1, **p<0.05, ***p<0.01.

Table 4. Horse Race Test: Tenure Status

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
SDR_Pub	0.00676***	0.00462***	0.00477***	0.00485***	0.00481***		
	(0.0007)	(0.0006)	(0.0006)	(0.0006)	(0.0006)		
WoS_Pub	0.00134**	-0.0004	0.0001	0.00048*	0.00047*		
	(0.0006)	(0.0003)	(0.0003)	(0.0003)	(0.0003)		
ln_SDR_Pub1						0.06339***	
						(0.0083)	
ln_WoS_Pub1						0.0083	
						(0.0050)	
ln_SDR_Pub							0.04509***
							(0.0074)
ln_WoS_Pub							0.0059
							(0.0047)
Zero_SDR_Pub=1							-0.05582**
							(0.0209)
Female		-0.04478***	-0.04962***	-0.04160***	-0.04488***	-0.03945***	-0.04000***
		(0.0110)	(0.0106)	(0.0097)	(0.0098)	(0.0093)	(0.0093)
Marital status		0.01561*	0.01448*	0.01368*	0.01653**	0.01681**	0.01720**
		(0.0079)	(0.0078)	(0.0077)	(0.0070)	(0.0071)	(0.0071)
Academic Age		0.07074***	0.07069***	0.06973***			
		(0.0034)	(0.0034)	(0.0034)			
Academic Age2		(0.00127***	(0.00126***	(0.00124***			
		(0.0001)	(0.0001)	(0.0001)			
Asian		0.05389*	0.05646**	0.06016**	0.05400**	0.05411**	0.05380**
		(0.0276)	(0.0258)	(0.0263)	(0.0264)	(0.0257)	(0.0257)
Black		0.08680***	0.07440**	0.07567***	0.07639***	0.08515***	0.08520***
		(0.0321)	(0.0292)	(0.0278)	(0.0275)	(0.0275)	(0.0276)
Hispanic		-0.0018	0.0055	0.0158	0.0120	0.0098	0.0105
		(0.0254)	(0.0235)	(0.0237)	(0.0236)	(0.0235)	(0.0235)
Others		0.0236	0.0274	0.0277	0.0248	0.0239	0.0237
		(0.0243)	(0.0227)	(0.0227)	(0.0230)	(0.0225)	(0.0226)
Field of Study FE (Coarse)	NO	NO	YES	NO	NO	NO	NO
Field of Study FE (Fine)	NO	NO	NO	YES	YES	YES	YES
Academic Age FE	NO	NO	NO	NO	YES	YES	YES
Observations	10366	9778	9778	9778	9778	9778	9778
R-squared	0.0347	0.4262	0.4443	0.4617	0.4761	0.4827	0.4824
Adjusted R-squared	0.0345	0.4256	0.4434	0.4565	0.4687	0.4754	0.4750

Notes: The dependent variable in this table is an indicator of tenure status in 2003. Each column is a separate regression, with different set of control variables. Column (1) include level of SDR publication count and level of WoS publication count. Column (2) add in demographic controls including, an indicator of female, an indicator of marital status, a linear trend of academic age, a square term of academic age, and race indicators. Column (3) add in field of study fixed effects by a coarse definition of fields of study (7 fields). Column (4) replace the field of study fixed effects from the one by the coarse definition to a fine definition (83 fields). Column (5) replace the linear and squared terms of academic age to a full set of academic age fixed effects. Column (5) is our preferred specification. Column (6) has the same specification as column (5) but replace level SDR and WoS publication counts with log(SDR+1) and log(WoS +1). Column (7) replace them with log(SDR), log(WoS), and an indicator for zero SDR. The standard errors are clustered at academic age level. *p<0.1, **p<0.05, ***p<0.01.

Table 5. Horse Race: by Quintiles of Academic Age

Junior → Senior	(1) Quintile = 1	(2) Quintile = 2	(3) Quintile = 3	(4) Quintile = 4	(5) Quintile = 5
Dep. Var = Salary					
SDR_Pub	0.00464*	0.00619***	0.00921***	0.00758***	0.01052***
	(0.00245)	(0.00234)	(0.00197)	(0.00177)	(0.00179)
WoS_Pub	0.00227	0.00279***	0.00318**	0.00303***	0.00354***
	(0.00148)	(0.00091)	(0.00129)	(0.001)	(0.00117)
Observations	2047	2175	1916	1986	1977
R-squared	0.133883	0.169215	0.123533	0.201415	0.151448
Adjusted R-squared	0.091709	0.126632	0.070635	0.150487	0.093655
Dep. Var = Funding					
SDR_Pub	0.02242*	0.03414**	0.04596***	0.02865***	0.02405***
	(0.01136)	(0.01544)	(0.00902)	(0.00555)	(0.00503)
WoS_Pub	0.00398	0.01846	0.01081	0.01099**	0.01378***
	(0.00677)	(0.01313)	(0.00666)	(0.00536)	(0.00348)
Observations	1294	1159	1206	1210	1190
R-squared	0.200916	0.18674	0.223501	0.189167	0.226101
Adjusted R-squared	0.143981	0.122316	0.159316	0.123169	0.159666
Dep. Var = Tenure					
SDR_Pub	-0.00073	0.00556***	0.00760***	0.00471***	0.00369***
	(0.00128)	(0.00148)	(0.00175)	(0.00108)	(0.00051)
WoS_Pub	-0.00002	-0.00088	0.00132	0.00066	0.00026
	(0.00065)	(0.00071)	(0.00114)	(0.00057)	(0.00041)
Observations	2025	1953	1992	1910	1898
R-squared	0.104062	0.238973	0.205691	0.127491	0.084093
Adjusted R-squared	0.063337	0.202617	0.166771	0.083313	0.035807

Note: This is the 2003 SDR sample. The ranges of each quintile of the academic age are: [1, 4], [5, 9], [10, 16], [17, 25], [26, 45].

Table 6. Horse Race: by Quintiles of Match Probability

	(1)	(2)	(3)	(4)	(5)
Match Quality from Low → High	**Quintile = 1**	**Quintile = 2**	**Quintile = 3**	**Quintile = 4**	**Quintile = 5**
Dep. Var = Salary					
SDR_Pub	0.01132***	0.00831***	0.00338*	0.00575***	0.00262
	(0.00262)	(0.00162)	(0.00177)	(0.00183)	(0.00165)
WoS_Pub	0.00210*	0.00155*	0.00419***	0.00361***	0.00594***
	(0.00123)	(0.00093)	(0.00134)	(0.00124)	(0.00139)
Observations	1668	1674	1665	1674	1678
R-squared	0.159493	0.241056	0.222262	0.229861	0.248865
Adjusted R-squared	0.088403	0.178179	0.158546	0.167134	0.186796
Dep. Var = Funding					
SDR_Pub	0.05155***	0.01756**	0.02433***	0.01554**	0.01768***
	(0.01753)	(0.00868)	(0.00544)	(0.00718)	(0.00621)
WoS_Pub	0.00373	0.01282**	0.00341	0.01709***	0.02011***
	(0.0095)	(0.00485)	(0.00276)	(0.00344)	(0.00331)
Observations	928	972	1055	1077	1096
R-squared	0.243907	0.195302	0.19187	0.261325	0.229858
Adjusted R-squared	0.123877	0.076405	0.083134	0.165111	0.130613
Dep. Var = Tenure					
SDR_Pub	0.00694***	0.00381**	0.00284***	0.00457***	0.00131
	(0.00105)	(0.00148)	(0.00092)	(0.00093)	(0.00088)
WoS_Pub	-0.00058	-0.00014	0.00102*	0.00013	0.00171***
	(0.00067)	(0.00073)	(0.00053)	(0.00073)	(0.00054)
Observations	1574	1617	1633	1648	1657
R-squared	0.464462	0.55502	0.528846	0.54065	0.567874
Adjusted R-squared	0.416215	0.516742	0.489427	0.502597	0.531981

Notes: This is the 2003 SDR sample. The ranges of each quintile of the match probability are: [0.5000, 0.6901], [0.6902, 0.7568], [0.7568, 0.8128], [0.8128, 0.8693], [0.8693, 1].

Table 7. Horse Race: by Match Rounds

Match Quality from Low → High	(1) **Quartile 4**	(2) **Quartile 3**	(3) **Quartile 2**	(4) **Quartile 1**
Dep.Var = Salary				
SDR_Pub	0.00601***	0.00588**	0.00576***	0.00660***
	(0.00217)	(0.00223)	(0.00158)	(0.0011)
WoS_Pub	0.00129	0.00234**	0.00168	0.00656***
	(0.0011)	(0.0011)	(0.00103)	(0.001)
Observations	1670	1665	1165	3859
R-squared	0.165848	0.232442	0.376459	0.184256
Adjusted R-squared	0.095974	0.167395	0.300094	0.155584
Dep.Var = Funding				
SDR_Pub	0.03476***	0.02116***	0.00792	0.02977***
	(0.00802)	(0.00603	(0.00657)	(0.00608)
WoS_Pub	0.00233	0.01282***	0.00568	0.01784***
	(0.00567)	(0.00479	(0.00458)	(0.00467)
Observations	1019	1006	709	2394
R-squared	0.228306	0.235426	0.250621	0.168647
Adjusted R-squared	0.120286	0.125829	0.093059	0.121278
Dep.Var = Tenure				
SDR_Pub	0.00262***	0.00548***	0.00357***	0.00357***
	(0.00095)	(0.00126	(0.001)	(0.0007)
WoS_Pub	0.00022	0.00062	0.00001	0.00117***
	(0.00038)	(0.00076	(0.00066)	(0.00036)
Observations	1607	1607	1142	3773
R-squared	0.500633	0.531406	0.580369	0.516813
Adjusted R-squared	0.457019	0.490479	0.527811	0.499429

Notes: This is the 2003 SDR sample. The ranges of the mean match rounds for the quartile 1-4 are: [0, 1], [1, 1.3076], [1.3076, 1.767], [1.768, 2]. The larger the mean match rounds, the lower the match quality. Note that we choose to use quartile instead of quintile of the mean math rounds is because that the first quartile has a large cluster at 1.

Table 8. Horse Race: by Quintiles of Name Frequency

	(1)	(2)	(3)	(4)	(5)
Match Quality from Low → High	**Quintile = 5**	**Quintile = 4**	**Quintile = 3**	**Quintile = 2**	**Quintile = 1**
Dep. Var = Salary					
SDR_Pub	0.00712***	0.00611***	0.00705***	0.01045***	0.01004**
	(0.00141)	(0.00179)	(0.00169)	(0.00335)	(0.00457)
WoS_Pub	0.00251**	0.00297***	0.00687***	0.00546	0.02006***
	(0.00104)	(0.00057)	(0.00162)	(0.00367)	(0.00639)
Observations	1940	1949	1920	1918	1985
R-squared	0.218619	0.211081	0.223103	0.191477	0.181616
Adjusted R-squared	0.162929	0.154202	0.168044	0.133144	0.123286
Dep. Var = Funding					
SDR_Pub	0.02747***	0.01906***	0.01390***	0.04056***	0.05406***
	(0.0062)	(0.00494)	(0.00471)	(0.0135)	(0.01584)
WoS_Pub	0.0051	0.00768***	0.02879***	0.04271***	0.07160**
	(0.00568)	(0.00249)	(0.00402)	(0.01365)	(0.0338)
Observations	1171	1182	1177	1184	1134
R-squared	0.241076	0.220955	0.266244	0.239645	0.200691
Adjusted R-squared	0.149482	0.127085	0.178976	0.147394	0.097991
Dep. Var = Tenure					
SDR_Pub	0.00333**	0.00313***	0.00533***	0.00634***	0.00830***
	(0.00133)	(0.00065)	(0.00086)	(0.00158)	(0.00205)
WoS_Pub	0.00056	0.0005	0.00091	0.00043	0.00432
	(0.00063)	(0.00032)	(0.00079)	(0.00161)	(0.00437)
Observations	1881	1900	1870	864	1902
R-squared	0.527515	0.522224	0.516449	0.488341	0.452809
Adjusted R-squared	0.492706	0.486823	0.481196	0.450276	0.411978

Notes: This is the 2003 SDR sample. The ranges of quintiles 1-5 of frequency of first name last name combination are: [1, 24], [25, 82], [83, 281], [282, 1423], [1423, 218422]. The larger the name frequency, the lower the potential match quality.

Table 9: Instrumental Variable Analysis

	(1)	(2)	(3)	(4)	(5)	(6)
Second Stage	**Salary**	**Salary**	**Funding**	**Funding**	**Tenure**	**Tenure**
SDR_Pub	0.0163***		0.0583***		0.00594***	
	(0.00181)		(0.00476)		(0.000746)	
WoS_Pub		0.0113***		0.0419***		0.00506***
		(0.00130)		(0.00410)		(0.000602)
N	10101	10101	6059	6059	9778	9778
R-sq	0.157	0.146	0.147	0.129	0.476	0.463
adj. R-sq	0.146	0.134	0.128	0.109	0.468	0.455
First Stage	**SDR_Pub**	**WoS_Pub**	**SDR_Pub**	**WoS_Pub**	**SDR_Pub**	**WoS_Pub**
SDR_Pub		1.0481		1.0270		1.0494
		(0.0469)		(0.0612)		(0.0471)
WoS_Pub	0.4165		0.4228		0.4178	
	(0.0193)		(0.0246)		(0.0198)	
F-stat of the First Stage	807.32	1762.56	36.35	938.95	863.08	1715.97

Notes: This is the 2003 SDR sample. This table shows the estimates of instrument variable analysis. The upper panel shows the estimates from the second stage and the lower panel shows the estimates form the first stage. Each column is a separate model, where column (1) and (2)'s outcome variable is ln(Salary), column (3) and (4)'s outcome variable is ln(Funding) and column (5) and (6)'s outcome variable is tenure status. Models in column (1), (3) and (5) are using WoS's publication counts to instrument SDR's and models in column (2), (4) and (6) are using SDR's publication counts to instrument WoS's.