NBER WORKING PAPER SERIES

DO TWO WRONGS MAKE A RIGHT?
MEASURING THE EFFECT OF PUBLICATIONS ON SCIENCE CAREERS

Donna K. Ginther
Carlos Zambrana
Patricia Oslund
Wan-Ying Chang

Working Paper 31844
http://www.nber.org/papers/w31844

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
November 2023

We thank the following Institute for Policy & Social Research staff who collected and analyzed the Gold Standard data used in this study: Abigail Bird, Haley Pederson, Quinn Maetzold, and Addison Lake; Genna Hurd managed the project. We thank Chris Bollinger and Shahnaz Parsaeian for comments on earlier drafts of this study. Any errors are our responsibility. All authors contributed to the study conception and design. Material preparation, data collection and analysis were performed by Wan-Ying Chang, Patricia Oslund, and Carlos Zambrana. The first draft of the manuscript was written by Donna Ginther and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript. Ginther, Oslund and Zambrana acknowledge financial support from SRI International Subcontract PO20663 to prime contract NSFDACS16C1234, T.O. 9. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Do Two Wrongs Make a Right? Measuring the Effect of Publications on Science Careers
Donna K. Ginther, Carlos Zambrana, Patricia Oslund, and Wan-Ying Chang
NBER Working Paper No. 31844
November 2023
JEL No. C26,J40,O30

ABSTRACT

This paper examines whether publication data matched to the Survey of Doctorate Recipients can be used for research purposes. We use Gold Standard data created to validate the publication match quality and compare these measures to publications assigned by a machine-learning algorithm developed by Thomson Reuters (now Clarivate). Our econometric model demonstrates that publications likely suffer from non-classical measurement error. Using horse race and instrumental variable models, we confirm that the Gold Standard data are relatively free from measurement error but show that the Clarivate data suffer from non-classical measurement error. We employ a variety of methods to adjust the Clarivate data for false negatives and false positives and demonstrate that with these adjustments the data produce estimates very similar to the Gold Standard. However, these adjustments are not as useful when publications are used as a dependent variable. We recommend using subsamples of the data that have better match quality when using the Clarivate data as a dependent variable.

Donna K. Ginther
Department of Economics
University of Kansas
333 Snow Hall
1460 Jayhawk Boulevard
Lawrence, KS 66045
and NBER
dginther@ku.edu

Carlos Zambrana
University of Kansas
zambrana@ku.edu

Patricia Oslund
Institute for Policy & Social Research
1541 Lilac Lane
Lawrence, KS 66045
poslund@ku.edu

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

Do two wrongs make a right? Measuring the effect of publications on science careers

Abstract: This paper examines whether publication data matched to the Survey of Doctorate Recipients can be used for research purposes. We use Gold Standard data created to validate the publication match quality and compare these measures to publications assigned by a machine-learning algorithm developed by Thomson Reuters (now Clarivate). Our econometric model demonstrates that publications likely suffer from non-classical measurement error. Using horse race and instrumental variable models, we confirm that the Gold Standard data are relatively free from measurement error but show that the Clarivate data suffer from non-classical measurement error. We employ a variety of methods to adjust the Clarivate data for false negatives and false positives and demonstrate that with these adjustments the data produce estimates very similar to the Gold Standard. However, these adjustments are not as useful when publications are used as a dependent variable. We recommend using subsamples of the data that have better match quality when using the Clarivate data as a dependent variable.

1. Introduction

The canonical labor market model posits that workers are paid their marginal product, but productivity is very difficult to measure for most workers. To do so would require detailed information on a worker's role in a firm and the output they produce. Productivity is easier to measure for PhD scientists because they produce publications that are publicly released and curated by organizations such *Web of Science* or *Scopus*. Data that links career outcomes such as research funding to scientific publications have flourished in the past decade because universities and funders are interested in tracking the output from research funding.[1] This paper analyzes the

[1] Projects like ORCID (https://orcid.org/) and *Web of Science* ResearcherID (https://www.researcherid.com/#rid-for-researchers) allow researchers to self-identify publications and in the case of ORCID affiliations and grants. Academic Analytics track faculty publications and citations at research universities (https://academicanalytics.com/). The iCite Portfolio analysis tool tracks publications and bibliometrics associated with NIH funding (https://icite.od.nih.gov/). Google Scholar (https://scholar.google.com/) allows users to create a research profile that lists publications, working papers and citations.

quality of *Web of Science* (WOS) publication data matched to the National Science Foundation's Survey of Doctorate Recipients during 2013-2016 using measurement error models and finds that with some adjustments these data can be used to model salaries and federal research funding.

The Survey of Doctorate Recipients (SDR) is a biennial longitudinal survey that started in 1973 and has collected data through 2019. The SDR tracks recipients of doctorates in science, engineering, or health from U.S. institutions. It is derived from a sample randomly selected from the Survey of Earned Doctorates (SED), a census of all research doctorates awarded by U.S. institutions.[2] The SDR contains detailed demographic and career information and limited data on publications and patents. The SDR is also used to report on the scientific workforce in the National Science Board's *Science and Engineering Indicators* (National Science Board, National Science Foundation 2020). The SDR is the longest-running survey of scientists and engineers in the U.S. Linking publications to SDR respondents would increase the value of the SDR to researchers and policymakers interested in studying the effect of publications on career outcomes as well as the demographic and career characteristics associated with the number of publications.

The problem of matching publications to individuals in the SDR is error-ridden. Consider the issue of matching a publication listed on a CV or Google Scholar to a publication database such as *Web of Science* (WOS). There are no standards for what should be listed on a CV or, for that matter, in a database like Google Scholar. Thus, these self-reported (or Google-reported) measures of publications may overstate what would likely appear in WOS. Alternatively, CVs and individual web pages may not be updated regularly and will underreport total publications.

[2] Both are maintained by the National Center for Science and Engineering Statistics within the National Science Foundation (NSF).

In addition, journals curated by WOS have changed significantly over time. Some refereed journals only appear in WOS several years after they started publishing. For example, *Contemporary Economic Policy* began publishing in 1982 under the name *Contemporary Policy Issues* but appears in WOS starting in 1997. Some refereed journals do not appear in WOS at all (e.g. *Economics Bulletin*). Thus, it is important to match the subset of publications from CVs that will appear in WOS.

A second approach to matching publications to researchers is to rely on self-reports. ORCID and Publons require individuals to identify their own publications. This process could also be measured with error because many researchers do not have an ORCID or ResearcherID on Publons. In addition, ORCID and Publons require the author to periodically update their records. Thus, like CVs, ORCID or ResearcherID publications may underreport total publications.

The SDR asked for self-reported publications in the 1995, 2001, 2003 and 2008 surveys, but only those appearing within the past five years of the survey year. Thus, even if a person accurately reported their refereed publications in the SDR, this measure will likely underreport total papers published by that person in refereed journals. Furthermore, individuals may "over-report" publications by counting publications that are not refereed or currently under review (Jiang, Chang, and Weinberg 2021).

Finally, there is no reason to believe that measurement error in publications matched to an individual researcher are random. Given a set of publications that could potentially be matched to an individual (e.g. download all publications associated with the name D. Ginther; the publication is either flagged as a true match or discarded from the set). Several researchers have shown that indicator variables and income in the Current Population Survey (CPS) are

subject to nonclassical measurement error (Aigner 1973; Bollinger 1996, 1998; Black, Berger, and Scott 2000). In the case of matching a single publication to a researcher, matching assignment can be considered an indicator variable. Let T be a (0,1) indicator that a single publication belongs to an author, X a (0,1) indicator that a publication is matched to the author (correctly or not), and $E = T\text{-}X$ the measurement error based on the assignment algorithm. If the matching algorithm erroneously assigns a publication to a researcher (T=0, X=1), then measurement error E = -1. If the matching algorithm fails to match a true publication (T=1, X=0), then the measurement error E = 1, otherwise it is 0. An individual's total publication record in the matching algorithm is the aggregate of a series of the assignments in X, with the measurement error in the author's publication count being the aggregate of E.

A second type of non-classical measurement error arises when individuals self-report a subset of their total publications as in the SDR. When the SDR asked for publication counts, it requested publications from the previous five years. In this case, the authors report w which is a subset of total publications x. Even if self-reported publications are entirely accurate, the variable w suffers from omitted variable bias (measurement error). Wolfe, Havemen, Ginther and An (1996) show that these "window variables" produce biased estimates. The obvious solution to the window problem is to use all publications observed instead of a "window" of publications measured at a specific point in time.

The argument for using the totality of an individual's publication record is supported by how research productivity is used to evaluate researchers. One recent request for a promotion evaluation asked: "Please provide a thorough and objective assessment of Dr. XXX's scholarship, especially on the significance of the work produced and its impact on the field." Another request for evaluation requires "A truly distinguished record of scholarship." These

types of evaluation requests happen whenever an academic is promoted with tenure or to full professor and asks the letter-writer to comment on the entire research portfolio. Thus, one common-sense approach to limiting measurement error is to use the entire publication record for an individual as opposed to a subset.

Given these challenges, it is clear that no public record of publications will be free from error. However, since research output is observable and important to researchers and policymakers, linking publications to scientists is an important endeavor. Our goal is to determine whether and how the publications matched to individual researchers in the SDR are useful for research purposes. In 2009 SRI contracted on behalf of the National Center for Science and Engineering Statistics (NCSES) with a company then known as Discovery Logic to match publications to the SDR.[3] Technical difficulties with the matching procedure as well as ownership changes at the company delayed the creation of the publication match. In 2018, NCSES contracted with the Institute of Policy & Social Research at the University of Kansas to create a second manual "Gold Standard" data set to evaluate the quality of the Clarivate data used in this study.

In what follows, we compare publication counts found by Clarivate's algorithm and the Gold Standard (GS) with the aim of establishing which of these methods is better at identifying the articles truly published by the 750 authors in our sample. Having two independent measures of publications allow us to use instrumental variable methods to address measurement error. These methods are described in most econometrics textbooks. Our paper is closely related Jiang, Chang, and Weinberg (2021) who used classical errors-in-variables econometric methods to

[3] Discovery Logic was purchased by ThomsonReuters. The intellectual property and science division of ThomsonReuters was spun off to create Clarivate Analytics in 2016.

assess how the Clarivate and SDR measures of publications are related to salaries. We take a somewhat different approach assuming that publications are subject to non-classical measurement error. After developing this econometric model, we examine how well the GS and Clarivate publication measures predict two self-reported career outcomes: salary and the likelihood of receiving federal research funding. Our goal is to determine whether the Clarivate data can be used to analyze career outcomes, and we find that with some adjustments the Clarivate data are useful.

2. Data

As mentioned previously, the SDR is a biennial longitudinal survey that tracks scientific career outcomes such as labor force status, employment sector, salary, the relationship of the job to PhD, whether an individual changed jobs and federal research funding. The SDR also contains detailed demographic information such as the respondent's age, race/ethnicity, marital status, foreign-born status, and field of degree at the time of PhD. In survey years 1995, 2001, 2003 and 2008 the SDR also included a question asking each respondent to provide a count of the number of articles they (co)authored in the previous 5 years.[4]

As part of the Institute for Policy & Social Research contract, the NSF created a stratified random sample of 750 respondents from the 1993-2013 waves of the SDR. The sample was stratified by sex, race/ethnicity, and average count of publications reported in the SDR. The sampling algorithm also controlled for degree year, broad doctorate field and name commonality. The stratified sample oversampled women and those employed in academia. The Institute for

[4] For example, in survey year 1995 the question asked for the number of articles respondents authored or coauthored that had been accepted for publication in a refereed professional journal since April 1990.

Policy & Social Research created a "Gold Standard" (GS)[5] dataset with hand-collected information on a person's publications from individual websites, CVs, and databases such as Google Scholar, affiliation information from these sources and LinkedIn, covariates from the SDR, and personally identifiable information provided by NSF to identify publications from Clarivate's *Web of Science* (WOS) database that match the names, fields, and affiliations of the 750 SDR respondents. Our approach began with a naïve set of name queries that downloaded all records that matched the last name and first initial of the SDR respondents. We then used the information gathered from manual searches for these individuals on the internet to identify publications associated with the SDR respondent. Appendix 1 provided details on data collection and our matching algorithm. Our Gold Standard algorithm identified 20,497 publications whereas the Clarivate algorithm identified 23,576 publications. Using the GS as benchmark, we found that overall precision was 0.762 and recall was 0.763 for the Clarivate match. This suggests that significant measurement error remains in the Clarivate data.

We restrict our analysis to comparisons between total publications identified by both the GS and Clarivate algorithms, and we keep the last survey year each author was observed in the labor force. Finally, even though our initial sample consisted of 750 authors, our outcomes of interest—salary and an indicator for receiving government support—can only be observed for individuals who were employed full-time. For this reason, we excluded 31 individuals who were out of the labor force in all survey years they participated.

[5] The Institute for Policy & Social Research "Gold Standard" data is a combination of manual searches and a data cleaning algorithm and is differs from the SRI manual collection of publications from sources besides WOS.

2.1 Descriptive Statistics

We provide information on the covariates in the sample in Table 1 which shows average total counts by demographic, educational, and employment groups in the sample, as well as for the entire SDR sample. The GS subsample is different than the SDR full sample. The GS sample has more women, non-whites, foreign-born and academics than the SDR sample. The distributions of fields of degree and of relation of job to field of degree, however, are very similar to that of the general population. We also show our outcomes of interest: salary and whether an individual reports receiving federal research funding (government support). We selected salary because we assume that the number of publications should be associated with higher salaries especially in academia. Salary was also used as an outcome in a related paper by Jiang, Chang, and Weinberg (2021). We selected government support because Ginther et al (2011, 2018) have shown that impact factor weighted publications increase the likelihood of receiving research funding from the National Institutes of Health (NIH).

Table 1: Descriptive Statistics of the Sample

Variable	**Gold Standard Subsample**		**SDR**		**Outcomes**	
	Mean	**N.**	**Mean**	**N.**	**Salary**	**Government Support**
Self-identified Gender						
Men	52%	371	69%	201,779	87,564	32%
Women	48%	348	31%	89,399	74,182	33%
Race/Ethnicity						
Hispanic	14%	101	6%	17,115	81,939	36%
White	44%	319	70%	203,402	83,501	35%
Black	12%	84	5%	15,119	81,054	22%
Asian	25%	183	18%	50,982	75,906	31%
Other	4%	32	2%	4,560	90,890	38%
Birthplace						
US-born	61%	461	74%	214,783	83,382	33%
Foreign-born	39%	258	26%	76,395	77,835	31%
Marital Status						

Not Married	29%	206	23%	67,164	71,311	31%
Married	71%	513	77%	224,014	85,440	33%
Name Commonality						
Common Name	41%	298	-	-	78,471	35%
Uncommon Name	59%	421	-	-	83,459	31%
Field of Doctorate						
Computer and Mathematical	7%	49	7%	20,275	80,424	27%
Biological & Agricultural	29%	212	27%	77,247	80,728	37%
Physical and Related	16%	112	18%	53,274	81509	38%
Social and Related	14%	102	13%	37,352	78124	17%
Psychology	12%	84	13%	36,953	71,645	29%
Engineering	15%	105	18%	51,172	87,878	38%
Health	8%	55	5%	14,901	93,138	33%
Computer and Mathematical	7%	49	7%	20,275	80,424	27%
Employer Sector						
4-year College/University	59%	424	46%	132,655	73,116	39%
Business, For-Profit	23%	162	36%	103,671	101869	23%
Other	19%	133	19%	54,852	82,831	23%
Relation of Job to PhD Field						
Closely related	71%	509	68%	196,944	83,668	36%
Somewhat related	23%	164	25%	72,453	78,275	29%
Not related	6%	46	7%	21,781	67,317	9%

We also show average outcomes for the Gold Standard sample by demographic characteristics. The higher earning respondents in the sample tend to be male, US-born, married, to have uncommon names, to have PhDs in engineering or health, to work in business (for-profit) occupations, and to hold jobs closely related to their field of degree. The incidence of government support in the sample does not vary significantly by gender, birthplace, marital status, or name commonality. The largest observed differences are between races, with Black respondents receiving much less support than respondents in other racial categories (22 percent vs. over 30 percent for the rest); fields of doctorate, where only 17 percent of those with degrees in social and related fields receive government support, versus around 30 percent or more in other fields; employer sector, with close to 40 percent of those in academia (versus 23 percent for those not in academia) receiving federal funding; and relation between job and field of PhD,

with at least 29 percent of those in jobs somewhat or closely related to their degree receiving government support, compared to only 9 percent of those in jobs not related to their degree.

Table 2: Average Total Publications

	GS		Clarivate		SDR		Pub. Match Probability	
Variable	**Mean**	**S.D.**	**Mean**	**S.D.**	**Mean**	**S.D.**		**N.**
All	17.2	30.3	17.6	33.2	11.4	23.6	77%	719
Self-identified Gender								
Men	16.6	25.7	17.2	28.7	11.8	22.8	78%	371
Women	17.9	34.6	18.0	37.5	11.0	24.5	77%	348
Race/Ethnicity								
Hispanic	17.0	26.0	18.0	26.7	11.7	18.0	78%	101
White	16.1	22.8	18.0	25.7	11.8	20.9	77%	319
Black	8.4	19.6	8.5	20.7	9.6	23.6	77%	84
Asian	22.0	41.0	20.1	46.7	10.4	28.4	77%	183
Other	25.1	49.6	21.9	47.7	16.9	33.7	78%	32
Birthplace								
US-born	16.4	29.2	18.0	33.5	12.8	26.4	77%	461
Foreign-born	18.7	32.3	16.8	32.8	8.9	17.6	79%	258
Marital Status								
Not Married	14.6	28.1	16.6	36.5	11.3	27.1	76%	206
Married	18.3	31.2	18.0	31.9	11.5	22.2	78%	513
Name Commonality								
Common Name	18.9	34.5	18.2	40.3	11.7	26.7	75%	298
Uncommon Name	16.0	26.9	17.2	27.2	11.2	21.3	78%	421
Field of Doctorate								
Computer and Mathematical	12.5	21.5	17.9	42.7	8.5	18.1	80%	49
Biological, Agricultural and Environmental	21.7	32.2	20.9	33.4	14.3	25.3	76%	212
Physical and Related	22.6	36.2	22.5	42.8	18.7	38.7	78%	112
Social and Related	5.7	11.2	7.3	15.5	5.6	9.9	76%	102

Psychology	12.6	22.5	13.8	23.8	8.0	13.2	78%	84
Engineering	17.8	37.1	15.4	30.8	6.4	11.4	77%	105
Health	20.4	31.1	23.7	38.0	13.7	24.1	79%	55
Employer Sector								
4-year College/University	18.3	33.4	18.4	34.7	11.5	25.3	78%	424
Business, For-Profit	14.8	24.3	14.1	30.7	8.9	18.6	75%	162
Other	16.7	26.3	19.4	31.6	14.1	23.5	75%	133
Relation of Main Job to PhD Field								
Closely related	20.2	34.3	20.7	37.5	13.1	26.1	78%	509
Somewhat related	10.3	14.3	11.1	18.5	7.7	16.2	75%	164
Not related	8.7	16.9	6.1	10.6	5.9	12.0	68%	46
Number of articles published since PhD graduation[*]								
0	22%		29%		41%			
1-10	39%		31%		31%			
11-20	15%		15%		12%			
21-40	13%		14%		8%			
41+	12%		12%		8%			

*Relative frequency out of 719 respondents.

Table 2 presents results of publication matches by these covariate characteristics. The Gold Standard and Clarivate identified very similar average publication counts per author, with counts being similarly distributed over groups defined by these characteristics. However, some differences are worth noting. The Gold Standard identified more average publications for Asian researchers (22 vs. 20.1) and in the Other racial category (25.1 vs. 21.9), but fewer for White researchers (16.1 vs. 18). Related to this, the GS also found more publications for foreign-born authors (18.7 vs. 16.8), but the opposite for US-born authors (16.4 vs. 18). Clarivate found more average publications for authors with a PhD in computer and mathematical sciences (12.5 vs. 17.9), social and related sciences (5.7 vs. 7.3), in health fields (20.4 vs. 23.7) and those working in occupations not related to their field of degree (8.7 vs. 6.1), but fewer for authors with a PhD in engineering (17.8 vs. 15.4). The SDR identified significantly lower counts than both the Gold

Standard and Clarivate (11.4 vs. 17.2 and 17.6). This is consistent across groups, with some exceptions where the differences are not statistically significant. These include average publications by Black researchers (9.6 vs. 8.4 and 8.5), researchers in physical and related sciences (18.7 vs. 22.6 and 22.5) and social and related sciences (5.6 vs. 5.7 and 7.3), and those working in occupations not related to their field of degree (5.9 vs. 8.7 and 6.1). This demonstrates that the SDR data are more likely to be undercounts of true publications and subject to the "window problem." In both the GS and Clarivate data, about 25% of the samples have more than twenty publications, indicating the skewness of the publication distribution.

Table 2 also includes average publication match probability, a variable provided by Clarivate that measures the likelihood that each publication is a true match to that author. Overall, these are also distributed evenly throughout these groups, but some categories stand out. Papers by authors in computer and mathematical fields had the highest average match probability (80 percent), while the articles with the lowest match probabilities were those of authors with common names (75 percent), working in business and other employment sectors (75 percent for both), and those in occupations somewhat related and not related to their field of degree (75 percent and 68 percent, respectively).

At the bottom of Table 2 we also report a distribution of respondents by groups of number of articles published, according to each publication measure. The Gold Standard finds that fewer authors have 0 publications than Clarivate (21.6 vs. 28.9), and the Gold Standard also finds more with 1 to 10 publications (38.5 vs. 30.7). However, both data sets agree that about 60 percent of authors in the sample have 10 or fewer publications. The SDR identifies a significantly higher share of authors with zero publications than the Gold Standard or Clarivate (40.8 vs. 21.6 and 28.9), which is consistent with the SDR undercounting publications.

Table 3 shows the correlation between the GS, Clarivate and SDR publications. The GS and Clarivate measures are highly correlated (r=.834). However, the SDR publication measure has a lower correlation with both the Clarivate and GS measures. This is to be expected since the SDR measure is aggregated over five years and only available in a limited number of survey waves.

Table 3: Correlation of Publication Measures

	Gold Standard	Clarivate	SDR
Gold Standard	1.000	0.834	0.651
Clarivate	0.834	1.000	0.693
SDR	0.651	0.693	1.000

2.2 False Positive and False Negative Publication Matches

Any process that links publications to individuals will be error-ridden. Figure 1 shows average publications by years since PhD. For most years since PhD, the GS and Clarivate find very similar counts. However, for individuals 6 years past PhD and 21 years past PhD, the GS finds significantly more publications, and in years 3 and 30 years past PhD, Clarivate finds more publications.

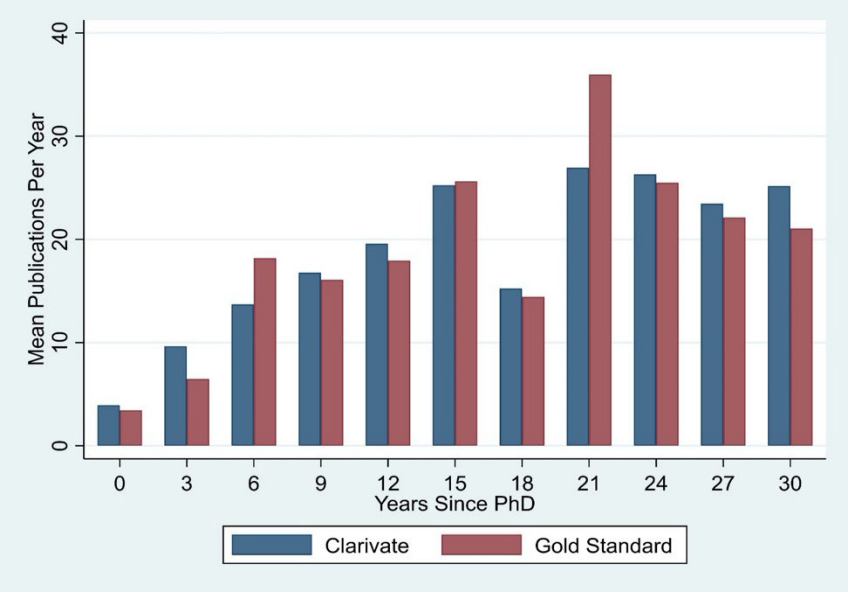

Figure 1. Mean of Gold Standard publications counts vs. Clarivate publication counts by years since PhD for Survey of Doctorate Recipient Respondents N=750.

For now, we assume that the GS approach is a close approximation to true publications because it used additional information sources to validate whether a publication should be attributed to an author.[6] Where Clarivate finds more publications than GS, we consider these excess publications to be false positives. Figure 2A shows a breakdown of publications by whether they were true or false positives. The percentage of false positive publications is higher for authors who graduated in the 2000s and 2010s (23 percent and 17 percent, respectively) when compared to those who graduated between 1960 and 1990 (between 10 percent and 13 percent), and it is lowest for those who graduated in the 1950s. In addition, the Clarivate and GS algorithm disagree on the number of respondents with zero publications. Clarivate reports 187 individuals with no publications compared to 132 in the GS data. We consider these records to be false negatives. Oslund, Ginther and Byrd (2020) found that 19.3 percent of Clarivate matches were false positives compared to the GS. They also found that 19.1 percent of publications were false negatives (where GS identified it as a publication).

We created variables to adjust for false negatives. We found that 64 SDR respondents reported having publications, but Clarivate identified zero publications for these people. False Negative-SDR is an indicator for false negatives when Clarivate reported zero publications, but the SDR reported positive publications. False Negatives-GS is an indicator for false negatives when Clarivate reported zero publications, but the GS reported positive publications. In Figure 2B we present a distribution of these false negatives by graduation decade. We can see that the proportion of false negatives is higher the earlier the graduation decades, reflecting the difficulty

[6] Analysis below confirms that the GS approach is relatively free from measurement error.

Figure 2A Figure 2B

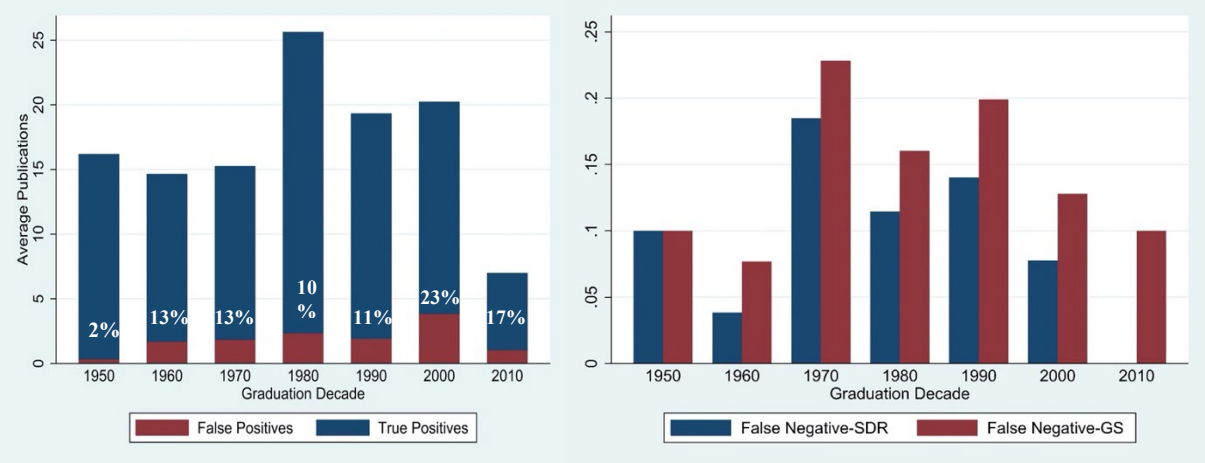

Figure 2a. Average Count and share of true and false positive publications by graduation decade in the Survey of Doctorate Recipients, N=750. Figure 2b Average percentage of publications that are false negatives in the Clarivate match compared to the Survey of Doctorate Recipients and Clarivate match compared to Gold Standard.

of finding matches for authors who may have published the bulk of their research before Researcher IDs existed. If they stopped producing research earlier in the 20th century, they may no longer be interested in identifying their research in publication aggregators such as Web of Science, thus making it more difficult for Clarivate to correctly identify their articles. The share of false negatives drops with each subsequent decade.

3. Empirical Strategy

Previously we argued that it is difficult (if not impossible) to identify a "true" measure of publications and that when we do measure publications, the process follows non-classical measurement error. To fix ideas, we estimate a model

$$y = \alpha + \beta x + \epsilon \quad (1)$$

where y is the outcome of interest – in this case salary or the probability of government support – and x is the true count of an individual's refereed publications. Our expectation is that $\beta > 0$; in

other words, that more publications increase one's salary and probability of receiving federal government support (e.g. grants). As discussed, publications in both the GS and Clarivate samples are likely measured with error. Let w^{GS} be the count of publications identified by the GS algorithm and w^{Clar} be the count of publications identified by the Clarivate matching algorithm. These measures are related to true publications x

$$w^{GS} = x + u^{GS}$$

$$w^{Clar} = x + u^{Clar}$$

Where u^{GS} and u^{Clar} are measurement errors.

3.1 Nonclassical Measurement Error

As we argued above the assignment of publications to individuals likely follows nonclassical measurement error, and the $Cov(x, u^{GS}) \neq 0$ and $Cov(x, u^{Clar}) \neq 0$. To focus on the measurement error issues, we abstract from other covariates that are present in the model. We assume that the $Cov(u^{GS}, u^{Clar}) = 0$. We do this because the algorithms used different pieces of information to identify whether a publication was a true match. The Clarivate method used self-reported ResearcherID publications as training data in a machine learning algorithm. The GS algorithm manually matched publications from CVs, Google Scholar, as well as field and affiliation to identify correct publications extracted from WOS. Thus, we are reasonably assured that the matching errors in the two algorithms are independent. To demonstrate the impact of nonclassical measurement error that is present in both the GS and Clarivate data, let w be the mismeasured publications (either from Clarivate or GS matches) and substitute in for the true value of $x = w - u$ into equation (1). The equation is now

$$\begin{aligned} y &= \alpha + \beta(w - u) + \epsilon \\ y &= \alpha + \beta w + (\epsilon - \beta u) \end{aligned} \tag{2}$$

Where the last term in equation is measurement error. After estimating equation (2),

$$\hat{\beta} = \frac{Cov(w, y)}{Var(w)} = \frac{Cov(x + u, \beta x + \epsilon)}{Var(x + u)}$$

We assume that the *Cov(u,ε)*=0 and the *Cov(x,ε)* = 0. Taking the probability limit,

$$plim(\hat{\beta}) = \frac{\beta(\sigma_x^2 + \sigma_{xu})}{\sigma_x^2 + \sigma_u^2 + 2\sigma_{xu}}$$

Unlike classical measurement error where the *Cov(x,u)* = 0, the bias in β depends on this quantity. β is biased towards zero as long as $(\sigma_x^2 + \sigma_{xu}) > 0$. In our case the *Cov(x,u)* > 0 when the algorithm undermatches true publications. Oslund, Ginther and Byrd (2020) show that the Clarivate data reported more people with zero publications than in the GS, which suggests that *Cov(x,u)* > 0 and β will be biased downwards.

We follow Jiang, Chang, and Weinberg (2021) in using "horse race" models to determine the comparative explanatory power of the GS and Clarivate matched data. The publication count measure with the least measurement error should also have the coefficient that is less biased and more likely to be statistically significant. Consider an expanded version of model (1):

$$y = \alpha + \beta x + \gamma c + \varepsilon \quad (3)$$

Where *x* measures publications and *c* are covariates such as demographic characteristics and PhD year. Since we do not observe *x*, we include both w^{GS} and w^{Clar} separately and together in the same model. We expect that the coefficients on the GS data will have less attenuation bias and will be more likely to be statistically significant. Although this is an informative comparison, other econometric approaches are required to determine the true effect of publications on our outcomes of salary and receiving research funding.

3.2 Instrumental Variables

Instrumental variables have been recommended to address the bias from measurement error. An instrument *z* is correlated with the mismeasured variable *x* but uncorrelated with the measurement error:

$$\hat{\beta}_{IV} = \frac{Cov(y,z)}{Cov(w,z)} = \frac{Cov(\beta x + \epsilon, z)}{Cov(x+u,z)}$$

Taking the probability limit,

$$\text{plim}(\hat{\beta}_{IV}) = \frac{\beta \sigma_{xz}}{\sigma_{xz} + \sigma_{zu}}$$

we can obtain consistent estimates as long as the instrument *z* is only correlated with the true variable *x*, but not with the measurement error u. However, that may not be the case in the presence of non-classical measurement error. Consider the case where the GS publication measure is an instrument for the Clarivate publication measure, $z = w^{GS} = x + u^{GS}$. In this case $\sigma_{zu}^{Clar} = \sigma_{xu}^{Clar} + \sigma_{u}^{Clar}\sigma_{u}^{GS}$ even when $\sigma_{u}^{Clar}\sigma_{u}^{GS} = 0$ because of the non-classical measurement error (Black, Berger and Scott 2000). If $\sigma_{xu}^{Clar} < 0$, then $\hat{\beta}_{IV}^{Clar} > \beta$ and the instrumental variables estimate will be biased upwards.

We estimate the IV models first by using GS publications as instruments for Clarivate matches and second by using Clarivate publications as instruments for GS matches. After obtaining the IV estimates we can use the Durbin-Wu-Hausman test for endogenous variables to evaluate the presence of measurement error in each IV regression. In these cases, the null hypothesis is that the publication measure being instrumented is exogenous. If we fail to reject the null, this indicates that the publication measure does not suffer from measurement error.

4. Results

Table 4 presents the results of running equation (3) for both outcomes. In the regressions we take the log of publications, $\log(w^j + 1), j \in \{GS, CLAR\}$, in order to estimate the effect of the

percentage change in publications on the log(salary) and the linear probability of receiving federal research support.[7] All estimates control for indicators for female, married, female×married, four categories of race/ethnicity (white being the omitted category), foreign-born status, and dummies for each survey year, together with dummies for zero publications for each of the counts. The estimates in Column (1) include the controls mentioned above, as well as a quadratic function of age. Column (2) adds indicators for 6 broad fields of degree, and employment sector. Column (3) – our preferred specification – swaps broad field for 21 indicators of narrow fields of degree and uses indicators of graduation decade instead of controlling for age.

Table 4: Horse Race GS, Clarivate and SDR

Panel A	**Log(Salary)**			**Government Support Indicator**		
Gold Standard	0.164***	0.165***	0.145***	0.077***	0.071***	0.067***
	(0.036)	(0.038)	(0.035)	(0.021)	(0.021)	(0.021)
Clarivate	-0.028	-0.011	-0.017	0.007	0.001	-0.003
	(0.034)	(0.035)	(0.034)	(0.020)	(0.020)	(0.019)
Observations	719	719	719	717	717	717
R-squared	0.166	0.200	0.283	0.110	0.142	0.209
Panel B						
Gold Standard	0.126***	0.122***	0.119***	0.097***	0.094***	0.107***
	(0.040)	(0.040)	(0.040)	(0.027)	(0.025)	(0.028)
Clarivate	-0.026	-0.014	-0.019	-0.037	-0.041*	-0.045*
	(0.034)	(0.032)	(0.032)	(0.025)	(0.023)	(0.024)
SDR	0.034	0.049	0.048	0.045*	0.031	0.022
	(0.036)	(0.036)	(0.036)	(0.025)	(0.024)	(0.025)
Observations	497	497	497	495	495	495
R-squared	0.248	0.293	0.348	0.095	0.170	0.216

[7] Since log salary is regressed on log publications, the coefficient on publications can be interpreted as a 1% increase in publications is associated with the coefficient estimate of percentage change in salaries. When the government support indicator is regressed on log salaries, the results are interpreted as a 1% increase in publications is associated with a .01 times the coefficient on publications change in salaries.

Covariates						
Gender	X	X	X	X	X	X
Marital Status	X	X	X	X	X	X
Female * Married	X	X	X	X	X	X
Race & Nativity	X	X	X	X	X	X
Survey Year	X	X	X	X	X	X
Age	X	X		X	X	
PhD Broad Field		X			X	
PhD Narrow Field			X			X
PhD Decade			X			X

Notes: All regressions estimated by OLS. Gold Standard = Log(GS Pubs + 1); Clarivate = Log(Clar Pubs + 1). We restrict observations to the last time observed in the labor force. Dependent variable in Columns (1)-(3) is the logarithm of yearly salary. Dependent variable in Columns (4)-(6) is an indicator for federal government support. All regressions control for gender and marital status fixed effects, as well as an indicator for female married, 4 race and nativity dummies (Black, Hispanic, Asian, Other), and 9 indicators for survey year. Columns (1), (2), (4) and (5) also control for a quadratic in age. Columns (2) and (5) include 6 dummies for broad field of degree (biological and agricultural sciences, physical sciences, social and related sciences, psychology, engineering, and health. The omitted category is computer and mathematical sciences.) Columns (3) and (6) replace dummies of broad fields for 21 dummies of narrow fields, and the quadratic in age for 6 indicators of PhD graduation decade. Standard errors clustered at the individual level (in parentheses). *** $p<0.01$, ** $p<0.05$, * $p<0.1$

The results in Table 4 Panel A show that GS publications have significantly stronger explanatory power than Clarivate publications in all specifications for both salary and government support, indicating they have a much greater signal. None of the Clarivate publication measures are statistically significant. These results are robust to changes in how narrow field of PhD is defined and how we control for experience (quadratic in age vs. dummies for PhD decade).

Given the predictive strength of SDR self-reported publications in Jiang, Chang, and Weinberg (2021), we compared Gold Standard and Clarivate counts with those measured in the SDR in Table 4 Panel B. Due to restrictions on the years for which the SDR collected

publications, we use data on articles published in years 1990-1995, 1998-2003 and 2003-2008 for all three publication measures. As in the previous table, GS publications are the best predictors of both salary and the likelihood of having government support, and this result is robust for both outcomes and across specifications. The lack of statistical significance of the SDR variables underscores how publication aggregates suffer from omitted variable bias consistent with the "window problem." Moreover, the coefficients for Clarivate counts are negative in all specifications, and the coefficient for SDR publications is significant only in the first specification for government support.

Taken together, the horse race models suggest that the GS publication measures have more explanatory power than either the Clarivate or SDR measures. Next we use IV methods to further explore this relationship.

4.1 Instrumental Variables Analysis

Table 5 presents the results of our instrumental variable analysis. Columns (1) and (4) of the top panel show $\widehat{\beta^{GS}}$ and $\widehat{\beta^{CLAR}}$, the average effect of publications on salary, estimated by OLS using Gold Standard and Clarivate publications, respectively. The OLS estimates indicate that a one percent increase in GS publications is associated with a 0.131 percent increase in yearly salary, but a one percent increase in Clarivate publications is associated with only an 0.085 percent increase in salary. This result suggests that Clarivate publications likely suffer from attenuation bias more than GS publications. Columns (2) and (5) show the IV estimates, $\widehat{\beta_{IV}^{GS}}$ and $\widehat{\beta_{IV}^{CLAR}}$, the corresponding coefficients estimated using two-stage least squares (2SLS). The GS IV estimate (0.121 percent) remains close to its OLS counterpart, indicating low measurement error in GS publications. The results for the probability of federal research support show the same pattern, the IV coefficient 0.063 is also close to its OLS counterpart 0.064. The Clarivate IV coefficients,

on the other hand, are higher than the GS IV coefficient for both salary and federal research support estimates, and almost twice the OLS coefficient in the salary estimates 0.152 percent vs. 0.085 percent. Furthermore, the Clarivate IV estimates are larger 0.152 for salary and 0.075 percent for federal research support than the GS IV estimates, 0.121 percent for salary and 0.063 for federal research support. This discrepancy suggests the presence of non-classical measurement error.

Table 5: Instrumental Variable Analysis

	Log(Salary)					
	(1)	(2)	(3)	(4)	(5)	(6)
	OLS	IV	First Stage	OLS	IV	First Stage
Gold Standard	0.131***	0.121***				0.858***
	(0.021)	(0.027)				(0.027)
Clarivate			0.705***	0.085***	0.152***	
			(0.022)	(0.020)	(0.024)	
Observations	719	719	719	719	719	719
R-squared	0.283	0.283	0.732	0.266	0.255	0.712
Endogeneity Test						
F-statistic		0.258			17.29	
p-value		0.612			0.000	
First stage F-statistic			54.97			69.50
	Government Support					
Gold Standard	0.064***	0.063***				0.857***
	(0.014)	(0.018)				(0.027)
Clarivate			0.704***	0.044***	0.075***	
			(0.022)	(0.013)	(0.016)	
Observations	717	717	717	717	717	717
R-squared	0.209	0.209	0.732	0.198	0.190	0.711
Endogeneity Test						
F-statistic		0.0237			9.779	
p-value		0.878			0.00184	
First stage F-statistic			54.88			69.06

*** p<0.01, ** p<0.05, * p<0.1

Notes: Gold Standard = Log(GS Pubs + 1); Clarivate = Log(Clar Pubs + 1). All regressions control for gender and marital status fixed effects, as well as an indicator for female married, 4 race and nativity dummies (Black, Hispanic, Asian, Other), 9 indicators for survey year, 21 dummies of narrow fields of degree, and 6 indicators of PhD graduation decade. Standard errors clustered at the individual level (in parentheses).

Given the difference in IV estimates, we performed tests for the null hypothesis that the endogenous publication measures are in fact exogenous and reported the results at the bottom of Columns 2 and 5. We fail to reject the null hypothesis that GS publications are an exogenous regressor in both the salary and government support estimates but reject the null that Clarivate publications are exogenous for both outcomes. Finally, coefficients from the first stage, in Columns (3) and (6), show that GS counts have a much higher signal-to-noise ratio than Clarivate's.

Together, these results bolster the case that the GS publication measure is close to the "true" measure of publications, and that the Clarivate publications suffer from non-classical measurement error. For the remainder of the paper, we consider the GS publication variable to be free from measurement error.

4.2 Alternatives to IV Estimates

Although IV approaches allow researchers to obtain unbiased estimates, in practice this approach is not feasible for the full sample of publications matched to the SDR. In the remainder of the paper, we evaluate alternative measures of Clarivate publications to determine whether the Clarivate publication measure can be used in statistical analysis despite the significant non-classical measurement error.

Clarivate's machine learning algorithm predicted the probability that a given publication was authored by a respondent and considered correct matches to be those with at least 0.5 probability. We next consider alternative measure of publications that limits the measure of

publications with a high match probability but not so high that it would harm the quality of any estimates obtained from it. This is an exploratory analysis that shows the trade-off between the signal and noise in the matched publications. In Table 6 we explore higher thresholds of match probability to evaluate how they affect the coefficient estimates and measurement error. Each coefficient comes from an OLS regression, using counts of publications above various thresholds. Higher thresholds result in less attenuated coefficients, indicating that those counts indeed have lower measurement error. The least attenuated coefficients in the salary equations are obtained when we keep publications with at least 0.80 and 0.85 match probability, with the former having a slightly lower standard error. These estimates are very similar to the GS estimates. Once the sample is limited to match probability of .9 or higher, the coefficient on Clarivate publications drops. Government support estimates also increase with publication match probability, but they do not reach a maximum around 0.80 or 0.85. Based on these results, we recommend that a threshold of 0.80 match probability was likely the best option. It is the conservative choice between 0.80 and 0.85.

Table 6: Estimates by Minimum Match Probability of Publications

	Log(Salary)						
	GS	**Clarivate**					
	All	All	P≥0.70	P≥0.75	P≥0.80	P≥0.85	P≥0.90
Log(Salary)	0.131***	0.085***	0.100***	0.108***	0.115***	0.115***	0.110***
	(0.021)	(0.020)	(0.020)	(0.021)	(0.022)	(0.024)	(0.028)
Observations	719	719	719	719	719	719	719
R-squared	0.283	0.266	0.270	0.273	0.274	0.271	0.264

*** p<0.01, ** p<0.05, * p<0.1

	Government Support Indicator						
	GS	Clarivate					
	All	All	P≥0.70	P≥0.75	P≥0.80	P≥0.85	P≥0.90
Government Support Indicator	0.064*** (0.014)	0.044*** (0.013)	0.050*** (0.014)	0.052*** (0.014)	0.061*** (0.015)	0.065*** (0.015)	0.077*** (0.018)
Observations	717	717	717	717	717	717	717
R-squared	0.209	0.198	0.200	0.201	0.206	0.206	0.208

*** p<0.01, ** p<0.05, * p<0.1

Notes: GS = Log(GS Pubs + 1); Clarivate = Log(Clarivate Pubs + 1). "All" refers to publication counts including all publications with match probability at least 0.5. The subsequent Columns use Clarivate publication counts including publications with at least 0.70, 0.75, 0.80, 0.85 and 0.90 match probability. All regressions control for gender and marital status fixed effects, as well as an indicator for female married, 4 race and nativity dummies (Black, Hispanic, Asian, Other), 9 indicators for survey year, 21 dummies of narrow fields of degree, and 6 indicators of PhD graduation decade. Standard errors clustered at the individual level (in parentheses).

Following Black, Berger and Scott (2000), we use a second approach to controlling for measurement error. They suggest that publications are less likely to be measured with error when both GS and Clarivate agree that publications are correctly assigned to authors. In Table 7 we estimate OLS models for our outcomes of interest on subsamples defined by different levels of agreement in publication counts. In Column (1) we limit the sample to those for which the GS and Clarivate algorithms agree on the number of publications. In Column (2) we allow the GS and Clarivate measures to differ by at most 3, in Column (3) we allow publications to differ by at most 5, in Column (4) we allow them to differ by at most 10, and in Column (5) we include the estimate that includes all publications. When observations where GS and Clarivate disagree are dropped from the specification, the coefficient on log publications is 0.115 (p<.01), as publications increase by 1 percent salary increases by 0.115 percent. The IV estimates in Table 5

are somewhat higher. This discrepancy is likely the result of losing 453 observations. However, it is important to note that this coefficient is identical to the one in Table 6, when we use Clarivate counts with at least .80 match probability. As we relax the degree of agreement, the coefficient falls, demonstrating attenuation bias from measurement error. For government support, the coefficient when there is no discrepancy between GS and Clarivate counts is 0.059, which is also closest to the corresponding coefficient in Table 6 when restricting publications to have at least 0.80 match probability.

Table 7: OLS by Count Difference between Clarivate and GS

	Log(Salary)				
	(1)	(2)	(3)	(4)	(5)
	0	0-3	0-5	0-10	All
Clarivate	0.115***	0.118***	0.110***	0.090***	0.085***
	(0.035)	(0.026)	(0.025)	(0.025)	(0.020)
Observations	266	474	534	605	719
R-squared	0.371	0.333	0.334	0.276	0.266

*** p<0.01, ** p<0.05, * p<0.1

	Government Support Indicator				
	(1)	(2)	(3)	(4)	(5)
	0	0-3	0-5	0-10	All
Clarivate	0.059**	0.053***	0.052***	0.046***	0.044***
	(0.027)	(0.018)	(0.017)	(0.016)	(0.013)
Observations	264	472	532	603	717
R-squared	0.284	0.233	0.224	0.209	0.198

*** p<0.01, ** p<0.05, * p<0.1

Notes: Clarivate = Log(Clarivate Pubs + 1). Column (1) restricts the sample to observations where Gold Standard counts and Clarivate counts are equal. Columns (2)-(4) add observations for which the absolute difference between Clarivate and Gold Standard counts is at most 3, 5 and 10 publications, respectively. All regressions control for gender and marital status fixed effects, as well as an indicator for female married, 4 race and nativity dummies (Black, Hispanic, Asian, Other), 9 indicators for survey year, 21 dummies of narrow fields of degree, and 6 indicators of PhD graduation decade. Standard errors clustered at the individual level (in parentheses).

4.3 Correcting for Clarivate Publications for False Positives and False Negatives

Based on the results in Table 6, we used Clarivate's match probability to correct for false positives by keeping only publications with at least 80 percent match probability (called Clarivate 80). We do so because this is the threshold that Oslund, Ginther, and Byrd (2020) found was associated with true positive matches in their analysis of match quality. Figure 3 shows a histogram of GS and Clarivate 80 publication counts by years since PhD. Throughout, GS finds more publications than Clarivate 80. Figure 3 shows the distribution of the number of articles published according to GS, Clarivate, Clarivate 80, and the SDR. GS finds fewer individuals with zero publications than the other measures. Clarivate 80 finds 100 more authors with zero articles than Clarivate and 153 more than the GS. This result is troubling because, in the interest of finding better measured counts, Clarivate 80 adds more false negatives to the already-large number in Clarivate counts (from 82 to 127).

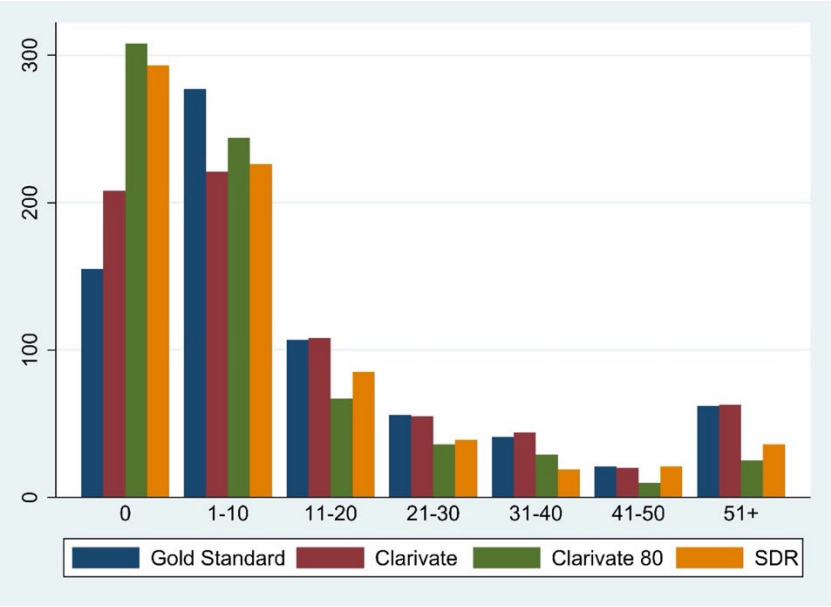

Figure 3. Distribution of number of articles published according to alternative publication measured of Gold Standard, Clarivate match, Clarivate match with .8 probability of correct match, and Survey of Doctorate Recipients self-reported publications.

Oslund, Ginther and Byrd (2020) found that Clarivate tended to under-match publications to individuals. They show that Clarivate has 55 additional respondents with no publications whereas the GS algorithm found publications. We consider including an indicator for false negatives equal to 1 for each observation where Clarivate found zero publications but the GS did. This approach, however, is not feasible for the full list of SDR respondents, so we also consider an analogous indicator constructed comparing Clarivate counts to SDR counts instead.

In Table 8 we use Clarivate 80 and corrections for false negatives in our estimates. All estimates are compared to GS. For reference, Columns (1) and (2) repeat the OLS estimates obtained using GS and Clarivate publications and presented in Table 6, and Column (3) repeats the estimates from restricting publications to those with at least 0.80 match probability (Clarivate 80). In Column (4) we use all Clarivate publications together with the false negative indicator based on SDR self-reported publications. The inclusion of this indicator increases the Clarivate coefficient from 0.085 to 0.11 for salary and 0.044 to 0.067 for government support, making both coefficients closer to the GS estimates. In Column (5) we use full Clarivate publications, but we drop observations with False Negative SDR equal to 1. This does not have a significant impact on the coefficient for salary, but the coefficient on government support increases to 0.074, which is higher than the GS OLS coefficient. In Column (6) we use Clarivate 80 publications and include a control for False Negative SDR. This effectively reduces the difference between the GS and Clarivate coefficients for salary estimates, and it increases the Clarivate coefficient for government support estimates even further to 0.077. Finally, in Column (7) we use Clarivate 80 publications and control for False Negative GS, which results in Clarivate coefficients still close to GS OLS for salary of 0.129 percent and reduces the coefficient for government support to

0.073. Overall, Clarivate 80 together with False Negative SDR in Column 6 provides the closest estimates to the GS OLS in Column 1.

Table 8: Estimates Correcting for False Negative and False Positives

Log(Salary)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
VARIABLES	Gold Standard	Clarivate	Clarivate 80	Clarivate	Clarivate	Clarivate 80	Clarivate 80
Publications	0.131***	0.085***	0.115***	0.110***	0.108***	0.131***	0.129***
	(0.021)	(0.020)	(0.022)	(0.022)	(0.022)	(0.023)	(0.023)
False Negative SDR				0.248**		0.196**	
				(0.104)		(0.099)	
False Negative GS							0.126
							(0.089)
Drop False Negative					X		
Observations	719	719	719	719	637	719	719
R-squared	0.283	0.266	0.274	0.273	0.284	0.279	0.277

*** p<0.01, ** p<0.05, * p<0.1

Government Support Indicator

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
VARIABLES	Gold Standard	Clarivate	Clarivate 80	Clarivate	Clarivate	Clarivate 80	Clarivate 80
Publications	0.064***	0.044***	0.061***	0.067***	0.074***	0.077***	0.073***
	(0.014)	(0.013)	(0.015)	(0.014)	(0.014)	(0.015)	(0.015)
False Negative SDR				0.227***		0.192***	
				(0.062)		(0.060)	
False Negative GS							0.104**
							(0.048)
Drop Bad Match					X		
Observations	717	717	717	717	635	717	717
R-squared	0.209	0.198	0.206	0.214	0.235	0.219	0.211

*** p<0.01, ** p<0.05, * p<0.1

Notes: Gold Standard = Log(GS Pubs + 1); Clarivate = Log(Clarivate Pubs + 1); Clarivate 80 = Log(Clar80 Pubs + 1), where Clar80 refers to Clarivate publication counts of publications with at least 0.80 match probability. All regressions control for gender and marital status fixed effects, as well as an indicator for female married, 4 race and nativity dummies (Black, Hispanic, Asian,

Other), 9 indicators for survey year, 21 dummies of narrow fields of degree, and 6 indicators of PhD graduation decade. Standard errors clustered at the individual level (in parentheses).

Table 9: Horse Race Models Correcting for False Negatives and False Positives

	(1)	(2)	(3)	(4)
VARIABLES	Log(Salary)	Log(Salary)	GovSup	GovSup
Gold Standard	0.108***	0.104***	0.045**	0.041*
	(0.037)	(0.037)	(0.022)	(0.021)
Clarivate 80	0.030	0.047	0.026	0.044**
	(0.039)	(0.041)	(0.022)	(0.022)
False Negative SDR		0.180*		0.186***
		(0.099)		(0.060)
Observations	719	719	717	717
R-squared	0.284	0.288	0.211	0.223

*** p<0.01, ** p<0.05, * p<0.1

Notes: Gold Standard = Log(GS Pubs + 1); Clarivate 80 = Log(Clar80 Pubs + 1), where Clar80 refers to Clarivate publication counts of publications with at least 0.80 match probability. All regressions control for gender and marital status fixed effects, as well as an indicator for female married, 4 race and nativity dummies (Black, Hispanic, Asian, Other), 9 indicators for survey year, 21 dummies of narrow fields of degree, and 6 indicators of PhD graduation decade. Standard errors clustered at the individual level (in parentheses).

To test how well Clarivate 80 explains our outcomes of interest when compared to the Gold Standard, we repeat the horse race analysis but this time using Clarivate 80 and a control for false negatives instead of the complete Clarivate measure of publications. The results of this analysis are presented in Table 9. Even though Clarivate 80 is a considerable improvement from full Clarivate, it still does not explain outcomes better than the Gold Standard.

In Table A4 in the appendix we present results for different subsamples of the data, comparing salary estimates using the Gold Standard, Clarivate, and Clarivate 80 that includes a correction for false negatives. The first panel presents OLS estimates using the full sample. We restrict the sample to academics, non-academics, individuals with common and uncommon

names, later years of the sample where Clarivate including full name with the publication, and people observed in the sample three times. Together these results suggest that the Clarivate 80 variables comes close to the GS estimates when we limit the analysis to subsamples. Next, we consider whether these adjusted variables can be used to as dependent variables.

4.4 Publication Measures as Dependent Variables

Researchers will also be interested in using publication measures as dependent variables. Measurement error in the dependent variable is less troublesome if there is classical measurement error but not with non-classical measurement error. To see why this the case, suppose as above that observed publications are $w = x + u$, where x is true publications and u is the measurement error, and that we would like to estimate the following model

$$x = \beta v + \epsilon$$

where ϵ is a random error term and *v* measures covariates. Since we do not know x, but we know that $x = w - u$, we can replace for x in the above equation to get

$$w - u = \beta v + \epsilon$$

Rearranging terms we get

$$w = \beta v + (\epsilon + u)$$

Because of non-classical measurement error, the $Cov(v, u) \neq 0$ but the $Cov(e, u) = 0$. Let $\hat{\beta} = \frac{COV(v,w)}{Var(v)} = \beta + \frac{Cov(v,u)}{Var(v)}$. This indicates that β will be biased either upwards or downwards depending on the sign of the $Cov(v, u)$.

Table 10: OLS Models of Publication Measures

	(1)	(2)	(3)
VARIABLES	Gold Standard	Clarivate	Clarivate 80
Black	-6.274**	-8.283***	-2.326
	(2.445)	(2.632)	(2.081)
Biology	10.467***	3.242	3.169
	(3.566)	(6.072)	(2.848)
Job somewhat related to PhD	-7.747***	-10.582***	-7.290***
	(2.888)	(2.388)	(1.678)
Test of equality with Gold Standard coefficient			
Black		0.872	-1.698*
Biology		2.461**	2.671***
Job somewhat related		-0.111	-1.685*
Observations	719	719	719
R-squared	0.208	0.179	0.191

Notes: Gold Standard = Log(GS Pubs + 1); Clarivate = Log(Clarivate Pubs + 1); Clarivate 80 = Log(Clar80 Pubs + 1), where Clar80 refers to Clarivate publication counts of publications with at least 0.80 match probability. All regressions control for gender and marital status fixed effects, as well as an indicator for female married, 4 race and nativity dummies (Black, Hispanic, Asian, Other), 9 indicators for survey year, 21 dummies of narrow fields of degree, and 6 indicators of PhD graduation decade. The Clarivate 80 model includes a control for false negatives measured in the SDR. The bottom panel presents test statistics and p-values for the test of equality between the coefficient of selected regressors in Columns (2) to (5) and their corresponding Gold Standard coefficient in Column (1). Standard errors clustered at the individual level (in parentheses).

In Table 10 we run models using three publication measures as dependent variables: The Gold Standard, Clarivate and Clarivate 80. We consider the GS to be the closest measure to "true" publications, but this measure is only available for the 750 individuals in our sample. Thus, we examine whether our alternative measures of publications provide similar results to GS. We selected three covariates where the GS estimates show a statistically significant relationship between the covariate and salary: Dummy variables for being a Black researcher,

having a PhD in biological sciences, and whether the job is somewhat related to the PhD. At the bottom of Table 10, we report the test statistic of the difference between each of three selected coefficients – dummies for Black, PhD in Biological Sciences, and for job somewhat related to PhD – and the corresponding one from the GS regression.

Using the GS variable we find that Black researchers publish an average of around 6.3 fewer publications than the omitted category of white researchers. Respondents who majored in biological sciences publish about 10.5 more papers than the omitted category, computer and mathematical sciences. The other models report coefficients significantly different from the GS regression. Finally, authors who considered their jobs to be in occupations somewhat different from their field of PhD publish around 7.8 fewer papers than those in fields closely related to their PhD. The coefficients from the Clarivate and Clarivate 80 are not statistically different from the coefficient in the GS regression for Black researchers and those whose job is somewhat related to their degree. The significant difference in the coefficient for biological sciences suggests that none of the alternative measures will result in reasonably unbiased coefficients when used for the full sample of SDR respondents. Taken together, measurement error still poses difficulties when using publications as a dependent variable in the full sample.

5. Conclusions

This study has documented the challenges of linking publication measures to individual scientists. This kind of match is difficult because no "true" measure of publications exists. We use hand-collected data on publications (Gold Standard) compared to a machine-learning algorithm (Clarivate) and measurement error models to determine whether the Clarivate publication measure can be useful in research. We demonstrated that the Clarivate matched data suffer from non-classical measurement error.

We estimate the effect of publications on salaries and the linear probability of receiving federal research (government) support using both publication measures. Using horse race models and instrumental variables estimates, we confirmed that the GS publication measures are relatively free from measurement error. We then examined whether correcting the Clarivate publication measures for false positive and false negative publications results in improved estimates. We found that controlling for false positives by keeping publications for match probability greater than or equal to 80 percent (the Clarivate 80 measure) and adjusting for false negatives by including an indicator results in estimated coefficients that are similar to the GS estimates for the full sample. We also found that these performed well for subsamples of the data including Academics, those with uncommon last names, publications from 2008-2013, and those observed at least three times in the SDR (results in the appendix). The Clarivate 80 publication measure performs less well when used as dependent variables. In that case, we recommend using Clarivate 80 variable with the academic, uncommon names or 2008-13 subsamples.

Although the Clarivate matched data suffer from non-classical measurement error, this research has demonstrated how the data can be adjusted in order to be useful for research purposes. Future research should consider how this measurement error affects longitudinal estimates in the SDR.

Our results have important implications for researchers and policymakers. Researchers need to understand that publications could potentially be measured with error and this measurement error will result in biased estimates of the effect of publications on outcomes. What is true of publications will also be true of citations and related bibliometrics. Research funders have increasingly emphasized the relationship between publications and research funding. Funders such as the NSF and NIH require that publications associated with grants be

acknowledged and reported to the agency. It remains an open question as to whether publications reported to agencies are accurate and timely. Finally, scientific fields are increasingly embracing pre-prints. Our results are limited to the subsample of publications curated by Web of Science. Future research should consider whether pre-prints have a similar impact on salaries and research funding.

6. Competing Interests

Donna Ginther is currently funded by the National Science Foundation grants SES-2152437 and SMA-1854849. Ginther, Patricia Oslund and Carlos Zambrana are funded by the National Science Foundation grant NCSE-2215606 to build upon the results of this paper. Wan-Ying Chang is an employee of the National Science Foundation.

7. References

Aigner, D. J. (1973), "Regression With a Binary Independent Variable Subject to Errors of Observation," Journal of Econometrics, 1, 49-60.

Black, D. A., Berger, M. C., and Scott, F. A. (2000), "Bounding Parameter Estimates With Non-Classical Measurement Error," Journal of Labor Economics, 15-507-28.

Bollinger, C. R. (1996), "Bounding Mean Regressions When a Binary Regressor is Mismeasured," Journal of Econometrics, 73, 387-399.

———(1998), "Measurement Error in the CPS: A Nonparametric Look," Journal of Labor Economics, 16, 576-594.

Ginther, D. K. Schaffer, W. T., Schnell, J., Masimore, B., Liu, F., Haak, L. and Kington, R (2011), "Race, Ethnicity, and NIH Research Awards," Science*,* 333(6045): 1015-1019.

Ginther, Donna K., Basner, J., Jensen, U., Schnell, J., Kington, R., and Schaffer, W. T (2018), "Publications as Predictors of Racial and Ethnic Differences in NIH Research Awards," PLoS ONE. https://doi.org/10.1371/journal.pone.0205929

Jiang X., Chang W.-Y., Weinberg B. A. (2021), "Man Versus Machine? Self-Reports Versus Algorithmic Measurement of Publications. PLoS ONE 16(9): e0257309. https://doi.org/10.1371/journal.pone.0257309

National Science Board, National Science Foundation (2020*,* "Science and Engineering Indicators 2020: The State of U.S. Science and Engineering, NSB-2020-1, Alexandria, VA: Author. Available at https://ncses.nsf.gov/pubs/nsb20201/.

Oslund, P., Ginther, D. K. and Byrd, A. (2020), "An Evaluation of the Quality of Publications Matched to the Survey of Doctorate Recipients," University of Kansas, Institute for Policy & Social Research.

Wolfe, B., Haveman, R., Ginther, D., An, C. B. (1996), "The "Window Problem" in Studies of Children's Attainments: A Methodological Exploration," Journal of the American Statistical Association, 91, 970-982.

Appendix 1: Data collection and Publication Matching Algorithm

This appendix summarizes the data collection and publication matching algorithm developed by the Institute for Policy & Social Research and described in Oslund, Ginther & Byrd (2020). Students searched the web by name for the participants. They used auxiliary information provided by the SDR such as PhD field, PhD institution, and employer affiliation to make sure that the correct person had been identified. Later in the process, information on publications from sources other than Web of Science was used to confirm the "ownership" of WOS publications. Similarly, additional information on employer affiliations gathered by the students was matched to WOS affiliation fields. Table A1 provides a summary of data sources, and Figure A1 shows how the sources contribute to the matching process.

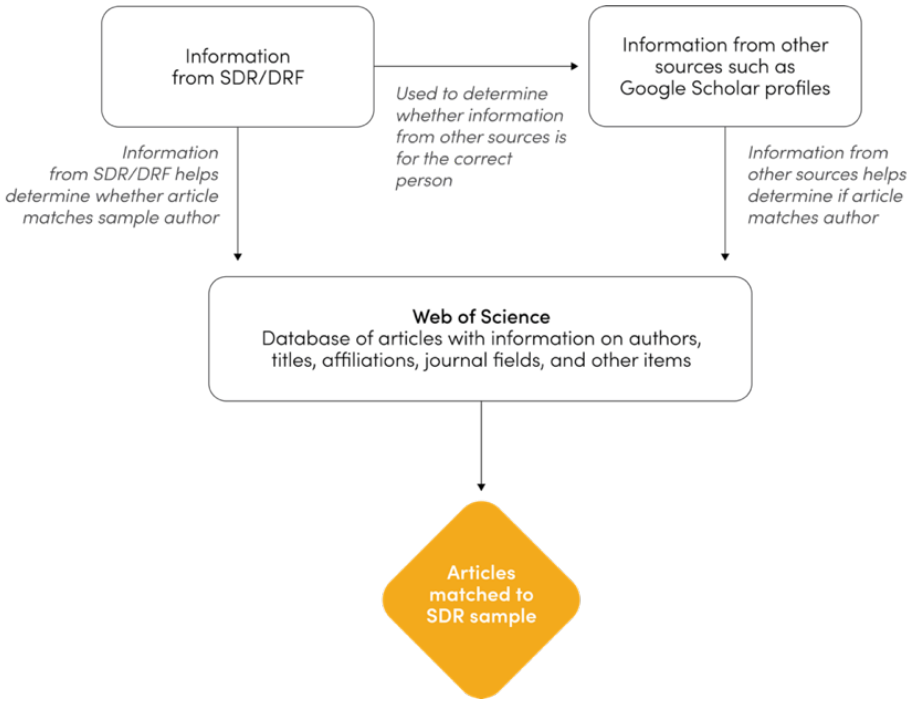

Figure A1. Flow of data for the publication matching process using data from the SDR, web searches and Web of Science.

Table A1. Summary of data sources

Source	Data Items	Comments
SDR-DRF (from NSF)		
	Name(s) Birth year PhD year, institution, field Employers-Affiliations Email addresses	The SDR is a sample survey of people receiving PhDs from US universities. The NSF surveys SDR recipients several times, approximately every other year. The DRF is a census of all PhD recipients from US universities, administered at the time the PhD is received. NSF provided us with a sample of 750 SDR participants, that is, 750 potential authors to be matched to publications.
University and personal web pages		
	Curricula vitae (CVs) Other publication lists Affiliations	We included information from web pages only when we were sure that the information belonged to the SDR participant and not another person with the same or similar name.
Google Scholar Profiles, Research Gate		
	Publications	Information from these sources was used only if we could verify that the information truly was for the SDR sample member.
ORCID/Researcher ID		
	Publications Affiliations ORCID and RID numbers	Both of these organizations assign unique alpha-numeric codes to researchers. Researchers are given the opportunity to identify their publications and to list their work history.
LinkedIn		
	Affiliations	Researchers with a LinkedIn presence generally list their PhD year, institution, and field. LinkedIn provides a reliable work history for cases where we can match the SDR participant's name and PhD information with a LinkedIn entry.
Web of Science		
	Publications	We extracted Web of Science information based on the SDR respondent's name. We used supplementary information from the other sources listed to decide whether the SDR participant was the true author of the article.

Details of the data collection process

The SRI Project search process consisted of two steps: first, general web searches and second, Web of Science (WOS) searches. For the general web searches, researchers used the individual's full name, degree, and affiliations (from SDR data) to find sources for additional information. Commonly found sources included CVs or university profile pages, LinkedIn profiles, and ResearchGate profiles. Less commonly found sources included Google Scholar, ORCID, and ResearcherID profiles. Researchers verified that sources found belonged to the correct individual and recorded information on affiliations and publications from them.

Researchers conducted WOS searches with the Advanced Search function and used constructed search strings, which include the individual's first name, last name, middle initial, and year restrictions. Organization restrictions based on affiliations were added if more than 1500 publications were returned (this threshold was later changed to 4500 publications returned).

Researchers faced three main challenges throughout the project. First, keeping case data within and outside the NORC Enclave caused issues with consistency and efficiency. Second, the SDR data that was used for the initial web search contained typos and errors in many cases. Third, certain situations made finding information on an individual during the general web search more difficult. These situations included an early PhD year, a common first or last name, an alternate first or last name, a name of Asian origin, a career outside of the United States, or a non-academic field or industry. For a detailed description of these situations, see Appendix 1 in Oslund, Ginther & Byrd (2020).

Overall, the project pulled together data from many sources with the goal of increasing the amount of information that can be used to decide whether a publication from Web of Science

matches a given author. We found Web of Science entries for 721 sample members using simple search strings based on name. However (as explained in later sections) some of the downloaded WOS records turn out not to be true matches. We found publication lists (other than WOS) for almost two-thirds of the sample, and LinkedIn entries for about half of the sample (Table A2). About 13 percent of the sample had publicly available ORCIDs, and about 5 percent had ResearcherIDs. These two identifiers provide a great deal of certainty when they are available, but unfortunately, they were not found for the bulk of the sample. The sample begins in 1993, more than 25 years ago. It is likely that many of the sample members left active research before LinkedIn, ORCID, and ResearcherID (RID) came into widespread use. Sample members who retired in the 1990s or early 2000s may never have placed CVs or other publications lists on the web. Information that might have been found on employer websites may have been removed once a sample member left employment.

Table A2. Number of sample members for whom data source was found.

Source	**Number of sample members found**	**Comments**
SDR-DRF (from NSF)	750	The original sample was provided by NSF.
CVs and other publication lists including Research Gate	471	In order to add data from a publication list, we needed confirming information that the information was for the correct person.

Google Scholar Profiles	107	Information was used only from people who provided public profiles with current institutional affiliation. The affiliation needed to match information from other sources. Information from these sources was used only if we could verify that the information truly was for the SDR sample member.
ORCID/Researcher ID	ORCID: 97 RID: 41	In addition to the ORCID or RID identifiers, the entries sometimes contained publication lists, which we included in the counts under CVs.
LinkedIn	378	We used this source to augment work histories.
Web of Science	721	For the most part, we downloaded articles using a match of either last name and first initial (if no middle initial) or last name plus 2 initials (if the article referenced two initials). When the number of records returned exceeded 4500 we added additional restrictions, most commonly an affiliation match. The WOS downloads contained many articles that we decided later were not a match.

Matching information from multiple sources

Once the Web of Science extracts were read into datasets, we created indicators of whether the WOS information matched information from other sources. A sequence of programs examined in sequence whether a publication matched a sample member on ten key elements:

1. Last name and first initial. We used alternative last name forms where information was available. For people with hyphenated last names, we generally used both parts of the compound names as alternatives. We did NOT include alternative spellings of last names (Hansen vs. Hanson) unless we encountered them in the SDR data or in our Web searches.
2. Last name and two initials. Asian names (such as Xiuying) often consist of two Chinese characters. We found that these names may be rendered as a single word or as two separate words (Xiu Ying), depending on the publication. The abbreviated name may use one or two initials (Wang, X or Wang, XY).
3. Last name and full first name. Continuing the example in item 2 above, we would look for Wang, Xiuying and Wang, Xiu.
4. Email address. Web of Science generally provided email addresses for the corresponding author only.
5. Publication title found in a CV or other publication list
6. ORCID
7. ResearcherID
8. Publication title found on Google Scholar Profile page
9. Academic or business affiliation
10. PhD field (matched to WOS journal field).

The last two elements in the matching list require additional explanation. Within WOS, an author's affiliation appears in an address field. Addresses may contain misspellings, abbreviations, and alternative renderings of institution names. For example, the University of Kansas may also appear as "University of KS," "Kansas University," "KU", and many other variants. Clarivate Web of Science provides lists of "preferred names" and "variant names" for most universities and large corporations. For every affiliation name found from Web searches or in the SDR, we searched for a WOS preferred name. If a preferred name was found, we searched all of the name variants to see whether any of them could be found in an address field. Sometimes a preferred name had over 1,000 variant names. If no preferred name was found (and hence, no variants), we reduced the affiliation name to one between or two key words and searched for that string within address fields. For example, if a sample member worked at "Lawrence Medical Associates," we would search for "Lawrence Med" in WOS address fields.

Mapping PhD fields to journal fields also presented challenges. We developed a mapping PhD field and likely journal fields of publication. Two sources contributed to the mapping: a measure of likelihood developed by Clarivate and provided to us by NSF, and a mapping created by Donna K. Ginther and Mabel Andalon (University of Melbourne). However, we note that over time a person's research interests may drift away from their original field—in fact, a person may be working completely outside of their field.

The KU matching algorithm

Our algorithm for matching publications to authors first categorizes articles based on information about author names. As mentioned above, some publications, especially those published before 2006, contain only author last names and first initials. Figure A2 below illustrates how we

classified name information into eight groups. When name information is limited, we require more supplementary information in order to determine a match.

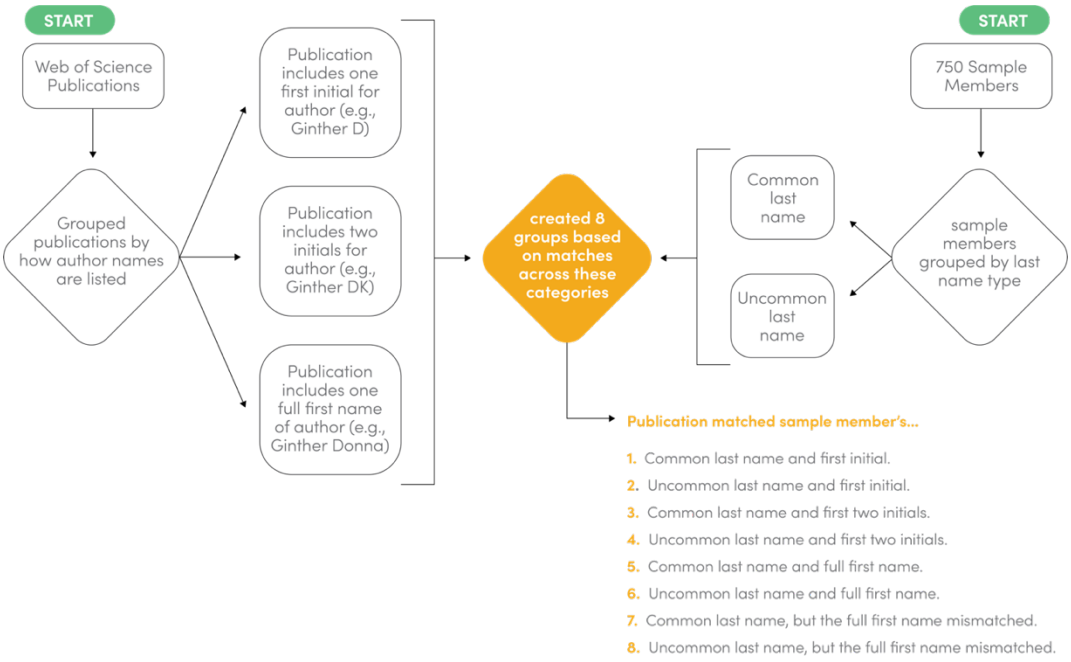

Figure A2. Categorizing information on author names starts with grouping publications by how author names are listed, groups based on matches and by common or uncommon name.

After sorting publications into categories, we examine other available information. Some information, when matched to the sample member, indicates that a match almost certainly is true. We have very high confidence in a match when:

- The Web of Science publication title matches a title found independently on a CV, Google Scholar profile, or other publication list.
- The email address on the WOS publication matches an email for the sample member.
- The ORCID or Researcher ID on the WOS publication matches the sample member.

Note that a first name mismatch ordinarily would cause us to reject that the publication was a match. However, a match on the factors listed above overrides that initial decision. We examined a few of these cases and found that they often were due to the use of a nickname for the publication author (Jim/James).

Other types of information that can lend certainty to a match include affiliation and field (Table A3). When both affiliation and field match the sample member, we flag the publication as a match. However, our confidence in the match is only moderate when a last name is common and only one initial is used on the publication. If only one piece of information matches the sample member (affiliation or field), our confidence in the match is low for cases with one or two initials and common last names. Our confidence is moderate when there is more name information. The use of full first names adds substantially to the confidence of our matches. After 2006 the use of full first names became much more common in Web of Science. Therefore, the perceived quality of our matches improves. Figure A3 traces the decisions employed by our matching algorithm for the case of a full first name and common last name.

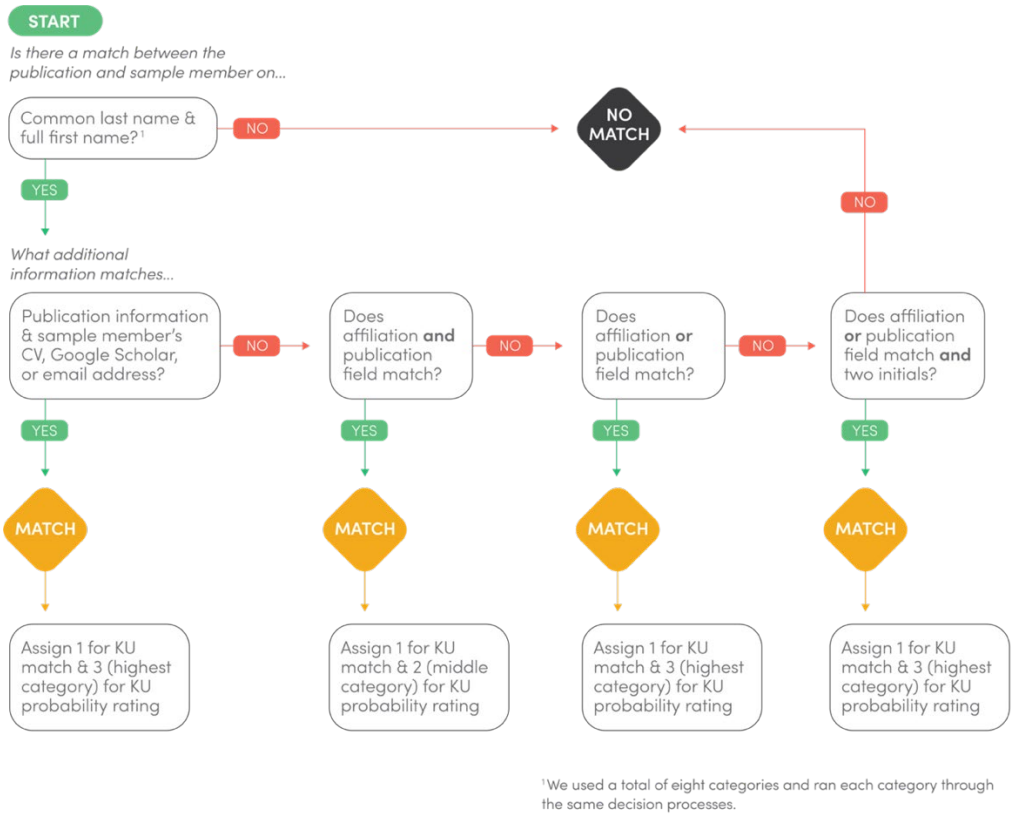

Figure A3. Example of the matching decision process based on the information we have from web searches, affiliations, common name, and first name or initials.

Table A3. Matching decision factors

Matching Group Based on Name	Decision Factors					KU Record Matches
	Match CV, GS, email, etc.	Else match Affil and Field	Else match Affil or Field	Else match (Affil or Field) plus 2 initials	Other	
Match one initial only, common last name	538	1,578				2,116
Match one initial only, uncommon last name	1,241	717	2,608			4,566
Match two initials, common last name	1,599	1,115	1,936			4,650
Match two initials, uncommon last name	2,104	965	1,034			4,103
Match full first name, common last name	1,473	404	696	189		2,762
Match full first name, uncommon last name	2,039	327	134	157	181	2,838
First name mismatch, common last name	42					42
First name mismatch, uncommon last name	17					17
Total	9,053	5,106	6,408	346	181	21,094

Light blue = 1. Low-certainty match. Medium blue = 2. Moderate-certainty match. Dark blue = 3. High-certainty match.

After running the matching algorithm, we create a data set of publications linked to sample members. Attached to each publication record is a broad measure of the certainty of the match. We use this measure of certainty as an indicator of the quality of the match and thus refer to match categories in terms of certainty or quality through the remainder of this report. We removed one sample member from the data because we found at a late stage that we were missing the person's affiliation. For the remaining 749 sample members, the algorithm identified:

- 12,927 type 3 (high-quality) matches;
- 3,623 type 2 (medium-quality) matches;
- 4,544 type 1 (low-quality) matches (Table 4).

For the purposes of this paper, we kept only high-quality (type 3) and medium-quality (type 2) matches in this paper.

Table A4: Subsample Analysis Log(Salary)

Publications	Gold Standard	Clarivate	Clarivate 80
Full Sample	0.131***	0.085***	0.131***
	(0.021)	(0.020)	(0.023)
Observations	719	719	719
R-squared	0.283	0.266	0.279
Academic	0.189***	0.126***	0.164***
	(0.026)	(0.024)	(0.026)
Observations	424	424	424
R-squared	0.385	0.352	0.362
Non-Academic	0.053	0.052	0.091*
	(0.036)	(0.035)	(0.046)
Observations	295	295	295
R-squared	0.284	0.285	0.295

Common Name	0.109***	0.054	0.100**
	(0.036)	(0.036)	(0.047)
Observations	298	298	298
R-squared	0.368	0.356	0.364
Uncommon Name	0.130***	0.098***	0.136***
	(0.028)	(0.027)	(0.029)
Observations	421	421	421
R-squared	0.328	0.313	0.329
Publications	Gold Standard	Clarivate	Clarivate 80
2008-2013	0.149***	0.131***	0.170***
	(0.029)	(0.026)	(0.031)
Observations	426	426	426
R-squared	0.312	0.302	0.318
Observed at least 3 times	0.115***	0.082***	0.128***
	(0.023)	(0.022)	(0.029)
Observations	439	439	439
R-squared	0.254	0.239	0.257

Notes: Clarivate 80 = Log(Clar80 Pubs + 1), where Clar80 refers to Clarivate publication counts of publications with at least 0.80 match probability. All regressions control for gender and marital status fixed effects, as well as an indicator for female married, 4 race and nativity dummies (Black, Hispanic, Asian, Other), 9 indicators for survey year, 21 dummies of narrow fields of degree, and 6 indicators of PhD graduation decade. Standard errors clustered at the individual level (in parentheses).

References

Oslund, P., Ginther, D. K. and Byrd, A. (2020), "An Evaluation of the Quality of Publications Matched to the Survey of Doctorate Recipients," University of Kansas, Institute for Policy & Social Research.