PRELIMINARY DRAFT

Addressing the Small-n Problem for the Scientific Workforce*

Holden A. Diethorn† Gerald R. Marschke‡ Elisabeth Ruth Perlman§
Joseph Staudt¶ Bruce A. Weinberg‖

January 16, 2024

Click here for the latest version.

Abstract

In this paper, we address the scientific workforce's "small-n problem" that has bedeviled prior work attempting to study the labor market outcomes of particular underrepresented groups. We do this by linking population-wide data on US-trained Ph.D. recipients from the Survey of Earned Doctorates (SED) to the universe of W2 and 1040 Schedule C (1040-C) tax records. This linkage allows us to track the entire US labor market history (between 2005-2020) for all Ph.D. recipients trained at US institutions for cohorts graduating between 2004 and 2015. In contrast to traditional sources of data, such as the Survey of Doctorate Recipients (SDR), this SED-tax linkage allows us to examine the outcomes of groups that have been traditionally too small to study. In particular, we not only examine earnings gaps separately by gender, race, and ethnicity, but also the gaps that exist when these characteristics are fully interacted. Moreover, our SED-tax sample of Ph.D. recipients allow us to examine how these earnings gaps vary across fields.

Keywords:
JEL Classifications:

*Any opinions and conclusions expressed herein are those of the authors and do not necessarily represent the views of the U.S. Census Bureau.

†Kenyon College. Email: diethorn1@kenyon.edu
‡SUNY Albany and NBER. Email: marschke@albany.edu
§U.S. Census Bureau. Email: elisabeth.perlman@census.gov
¶U.S. Census Bureau. Email: joseph.staudt@census.gov
‖Ohio State University and NBER. Email: weinberg.27@osu.edu

Timeline For Results

- Jan 17 - Jan 30: Finalize samples defined by the strength of Ph.D. recipient attachment to the US labor market. For instance, in order to be in our main sample, should a Ph.D. recipient be required to have positive earnings one year after graduation or *all* five years after graduation. This work is mostly finished.
- Jan 31 - Feb 14: Generate earnings gap estimates by sex, race, ethnicity and the interactions of all three characteristics. Also generate breakdowns by field. This work is mostly finished.
- Feb 15 - Feb 29: Prepare disclosure request and submit for review by Census.
- Mar 1 - Mar 15: Census reviews output to be disclosed.
- Mar 16 - Mar 30: Incorporate results into paper.

> "...the grand challenge is to solve the small-n problem, whereby underrepresented groups are accurately counted and their intersections of identity are given voice and visibility, while simultaneously ensuring the protection of individual privacy...a major impedance to developing actionable and effective policy interventions that are designed to improve equity for underrepresented groups is a lack of comprehensive and visible intersectional data."— Kaye Husbands Fealing and Aubrey DeVeny Incorvaia, *Understanding Diversity: Overcoming the Small-n Problem*, 2022.

1 Introduction

Earnings gaps across demographic groups that comprise the scientific workforce, including those defined by gender, race, and ethnicity, have been extensively studied.[1] However, the "Small-n Problem" (Fealing and Incorvaia, 2022) has limited progress on understanding the labor market outcomes of some underrepresented groups, especially when examined in combination with other demographic characteristics. This challenge becomes all the more bedeviling when trying to understand how these group differences vary across fields.

In this paper, we overcome the constraints typically imposed by small sample sizes for particular demographic groups by linking population-wide data on Ph.D. recipients from the Survey of Earned Doctorates (SED) to the universe of W2 and 1040 Schedule C (1040-C) tax records. This linkage allows us to track the entire US labor market history (between 2005-2020) for all Ph.D. recipients trained at US institutions for cohorts graduating between 2004 and 2015. In contrast to traditional sources of data, such as the Survey of Doctorate Recipients (SDR), this SED-tax linkage allows us to examine the outcomes of groups that have been traditionally too small to study. In particular, we not only examine earnings gaps separately by gender, race, and ethnicity, but also the gaps that exist when these characteristics are fully interacted, which, to our knowledge, has never before been feasible. Moreover, our SED-tax sample of Ph.D. recipients allow us to conduct a detailed examination of how these earnings gaps vary across fields.

We first examine gender earnings gaps. Conditional only on graduation year, we find that female Ph.D. recipients earn XX% more/less than their male peers, which increases/decreases

[1] For instance, on gender earnings gaps see work by Ceci et al. (2014); Li and Koedel (2017); Foster, McEntarfer, and Sandler (2022); Ceci, Kahn, and Williams (2023)... On earnings gaps by race, see work by Li and Koedel (2017); Foster, McEntarfer, and Sandler (2022)... On gaps by ethnicity, see... On the intersection of these characteristics, see...

to XX% five years after graduation. Adding covariates controlling for field, university, sector of employment (i.e. whether they work at a university or not), and demographic characteristics, the gap is attenuated/expanded to XX% and XX% at one and five years after graduation. Thus, even conditional on a variety of covariates... These results are in line / out of line with previous work on the gender earnings gap among Ph.D. recipients, including ADD CITATIONS HERE. A Gelbach decomposition suggest that most of the explained log earnings gap is accounted for by LIST COVARIATES. A sensitivity analysis reveals that an unobserved covariate such as productivity, would need to explain a lot/little of the residual variation in both the log earnings and the gender indicator to bring the gap to zero.

Turning to earnings gaps by ethnicity, we find that, conditional on graduation year, Hispanic Ph.D. recipients initially (one year out) earn XX% more/less than their non-Hispanic counterparts, with this gap increasing/decreasing to XX% after five years. Adding covariates increases/decreases these gaps and Gelbach decompositions suggest that the covariates most responsible for this attenuation/expansion are LIST COVARIATES. SENSITIVITY. COMPARISON TO OTHER WORK.

We next examine gaps between six race groups, finding that, conditional on graduation year, RACE1 and RACE2 Ph.D. recipients initially earn XX% and XX% more than their white peers, while RACE3, RACE4, and RACE4 initially earn XX%, XX%, and XX% less. AFTER 5 YEARS, ... ADDING COVARIATES... GELBACH DECOMPOSITION AND SENSITIVITY...COMPARISON TO OTHER WORK.

As noted, gaps across demographic groups of Ph.D. recipients have been well-studied. However, the "Small-n Problem" is particularly thorny when trying to examine these groups in combination. A primary strength of our comprehensive SED-tax linked data is that we can study the interactions between these groups. Combining the two gender categories, two ethnicity categories, and six race categories, we examine earnings gaps between the $2 \times 2 \times 6 = 24$ fully interacted demographic groups, which, to our knowledge, has never before been feasible.

We find that, relative to non-Hispanic white male Ph.D. recipients, ... FILL In RESULTS.

Finally, we turn heterogeneity in these gaps across field. FILL In RESULTS.

This analysis serves two main purposes. First, it introduces new SED-tax data infrastructure that links the universe of US-trained Ph.D. recipients to comprehensive US tax

records, allowing unprecedented opportunities to study this portion of the scientific workforce. Second, it illustrates the types of analyses that can be conducted with these data, by presenting earnings gaps for detailed demographic groups and further breaking them out by field of study. In future work, we aim to link Ph.D. recipients in the SED to their employers in order to better understand the characteristics of the jobs they are taking and of the firms that hire them.

1.1 Literature Search

- Stansbury and Schultz (2023) uses SED data to study socioeconomic diversity of Econ PhDs — despite being focused on Econ profession, the tables and figures show different measures of socioeconomic diversity across many fields (JEP article; Working paper version). Figure 3 in JEP article looks at female-share and URM-share across fields. (Kniffin (2007) another article showing differences in propensity to pursue a PhD based on parental education).
- Foster, McEntarfer, and Sandler (2022) looks at diversity and labor market outcomes in the economics profession using SED-LEHD data. Contains Mincer regressions.
- Ceci, Kahn, and Williams (2023) look at gender bias across six different domains in academic science. Interesting preface and intro. Look at salary gaps — finding similar to Li and Koedel (2017) — which they seemingly identify as the singularly strong article that tries to explain sources of gender gap. Compares findings to the raw gender gap in AAUP Faculty Compensation Survey — finds similar raw gap in SDR, but explained by various factors.
  - "Additionally, we will not address intersectionalities of gender with race and social class because very limited data exist."
  - "The literature on academic STEM salary gaps... is sparse and tends to be based on individual institutions and organizations. Moreover, even when national data re reported, they are limited by lack of controls for potentially confounding factors."
- Li and Koedel (2017) look at salary gaps using data they collected from 40 selective public universities. "We decompose racial-ethnic and gender wage gaps and show that academic field, experience, and research productivity account for most or all of the gaps. We find no evidence of wage premiums for individuals who improve diversity,

although for Black faculty we cannot rule out a modest premium."

- Ceci et al. (2014) shows (in Table 4 and Figure 17) gender earnings by field at different academic ranks (Asst Prof, Assoc Prof, Full Prof) using SDR 2010. Table A1 gives some measure of productivity by field (pubs conferences, patent apps, patents granted).
- Fealing and Incorvaia (2022) and the "small-n" problem.
- Bayer and Rouse (2016): Studies diversity in economics profession. "We then review evidence on how diversity affects productivity..."
- Lundberg and Stearns (2019) gives evidence that increases in female representation that occurred in 1970s and 1980s is now "stalled" in the economic profession.
- Ginther, Kahn, and Schaffer (2016) find that "race/ethnicity explain the NIH funding gap for women of color, as white women have a slight advantage over men in receiving Type 1 awards."
- Ginther and Kahn (2013) explore the leaky academic pipeline for women of color.
- Ginther et al. (2011) is an article in *Science* using the NIH IMPAC II grant database to look at NIH funding by race.
- NCSES report on "Diversity and STEM: Women, Minorities, and Persons with Disabilities" (`https://ncses.nsf.gov/pubs/nsf23315/`). The report is not focused on PhDs, but Table 2-4 gives PhD degrees by field, citizenship, race, and ethnicity b/w 2011–2020.
- Sowell, Allum, and Okahana (2015) studies minority PhDs (`https://legacy.cgsnet.org/ckfinder/userfiles/files/Doctoral_Initiative_on_Minority_Attrition_and_Completion_2015.pdf`)
- Webber (2015) uses SDR to look at differences in salary by gender and race.
- Minority representation in Econ: `https://www.aeaweb.org/content/file?id=426`; some info on BA and MA level as well.

2 Data

As discussed in the introduction, a key contribution of this paper is the use of population-wide data on Ph.D. recipients from the Survey of Earned Doctorates (SED) linked to the universe

of W2 and 1040 Schedule C (1040-C) tax records, which enables us to comprehensively track the domestic labor market outcomes of subgroups that have historically been too small to study using traditional sources such as the Survey of Doctorate Recipients (SDR). In this section, we provide a detailed overview of these comprehensive data sources used as well as the links between them. We also discuss how the SED-tax linked data compares to data available in the SDR.

2.1 Survey of Earned Doctorates (SED)

Survey of Earned Doctorates (SED) The NSF's Survey of Earned Doctorates (SED) is a census of US-trained doctorates conducted in the year of Ph.D. graduation. The SED asks respondents to provide their demographic characteristics including sex, race and ethnicity, age, parent's education, marital status, and family size. SED respondents also provide background educational characteristics including their fine field of study, date of Ph.D. graduation, the US institution where they received their Ph.D. degree, their source of primary financial support during Ph.D. training, and their post-PhD employment plans (e.g., whether they are planning to pursue a postdoc, whether they have a job offer). A key strength of the SED is its consistently-high response rates which has exceeded 90% every year between 1970 and 2022 (National Center for Science and Engineering Statistics, 2023).[2] Some shortcomings of the SED are that 1) it does not capture information from doctorates who receive their Ph.D. outside the US and 2) it only captures information at a single point in time.

Survey of Doctorate Recipients (SDR) The NSF's Survey of Doctorate Recipients (SDR) is a biennial survey of US-trained doctorates who received a Ph.D. in a STEM field.[3] Some key strengths of the SDR is that 1) it allows for longitudinal analyses of a sample of SED respondents, including those who leave the US after graduation, 2) it includes information about PhD job characteristics including self-reported annual salary, whether the doctorate is employed in academia, the academic rank of the doctorate (e.g., whether the PhD is employed as a postdoc, a non-tenure-track academic position, a tenure-track position, or has already obtained tenure).[4] A few shortcomings of the SDR are 1) that it does not cover Ph.D. recipients from non-STEM fields, 2) sample size limitations make it difficult to study

[2]SED response rates (1970-2022): https://ncses.nsf.gov/pubs/nsf24300/table/A-2.

[3]SDR respondents are sampled from the SED and as of 2021 represent an approximate 10% sample of the target population — US-trained STEM Ph.D. recipients under 76 years of age (https://ncses.nsf.gov/surveys/doctorate-recipients/2021#methodology).

[4]The SDR began tracking US-trained Ph.D. recipients who live outside the US as of 2010.

underrepresented minorities in STEM fields, and 3) lacks employer information needed to conduct extensive analyses of economic outcomes.[5]

Linking SED to Census Data For our analysis, we start with the 2000-2019 SED waves, which are currently available for analysis at the U.S. Census Bureau. These SED respondents have been transferred to the U.S. Census Bureau, which uses personally identifiable information (PII), such as name and birth date, to assign each Ph.D. recipient to an internal person-level identifier called a Protected Identification Key (PIK).[6] Once a Ph.D. recipient is assigned to a PIK, they can be linked to a wide array of confidential data at Census. In particular, as we discuss below, we can link them to the universe of W2 and 1040 Schedule C (1040-C) tax records.[7] In addition, we link them to comprehensive administrative information on their demographic characteristics (in particular, race, ethnicity, and gender), which supplements the self-reported information in the SED.

2.2 Individual Characteristics File (ICF)

Though self-reports from the SED are our primary source of demographic information about Ph.D. recipients, we augment this data with demographic information available in survey and administrative records from the Social Security Administration (SSA) and the U.S. Census Bureau. Specifically, we use the demographic information available from the Individual Characteristics File (ICF), a key component of the Longitudinal Employer Household Dynamics (LEHD) program's data infrastructure at Census (Vilhuber, McKinney, et al., 2014). SED respondents are linked to the ICF via their PIK.

Since we are primarily interested in estimating earnings gaps by gender, race, and ethnicity, we focus on these variables from the ICF. The sex variable in the ICF is composed of two values: "Female" and "Male" and the ethnicity variable is also composed of two values "Hispanic" and "non-Hispanic". The race variable takes on the following six values "White", "Black", "American Indian/Alaska Native", "Asian", "Native Hawaiian/Pacific Islander", and "Two or more races"

[5]The SDR includes a limited set of employer characteristics that are provided by the respondent such as their employer's size and whether their employer is a new business founded in the last five years.

[6]PIK assignment is done through the Person Identification Validation System (PVS), which is a probabilistic match (Wagner, Lane, et al., 2014)

[7]Of the XX respondents in the SED, XX% are assigned at least one PIK, XX% are assigned a unique PIK, and XX% are assigned a unique PIK that is not assigned to any other SED respondent (i.e., there is a 1-to-1 mapping). Our primary analysis focuses on this 1-to-1 sample.

Combining SED and ICF Demographic Characteristics As noted, the race and ethnicity categories in the SED are quite detailed, but they can be directly mapped to the six race categories and two ethnicity categories available in the ICF. We give the self-reported SED gender, race, and ethnicity values priority, and if the SED has a missing value, then it is filled using the ICF. The rate of missingness in the SED is XX% for race and XX% for gender. After augmentation with the ICF, these rates fall to XX% and XX%.

2.3 Earnings Records

SED respondents with an assigned PIK are linked to two sources of administrative earnings data: 1) IRS W2 tax records and 2) IRS 1040 Schedule C (1040-C) tax records. Taken together, these provide a complete picture of the domestic earnings history of each SED respondent between 2005 and 2020, including earnings from universities, private firms, government, and self-employment.

The W2 tax records provide annual information on individual earnings along with a federal tax identification number (EIN). As the name suggests, 1040-C tax records are derived from 1040 tax forms with attached schedule C(s) and are available through the U.S. Census Bureau's Integrated Longitudinal Business Database (ILBD) (Goetz, Kroff, et al., 2021). The ILBD includes the universe of non-employer firms in the United States and can be used to measure earnings from self-employment.[8] Overall, non-employer businesses account for about 75% of businesses in the United States (Jarmin, 2007), so incorporating income derived from self-employment is crucial to properly measuring faculty commercial engagement.

Total earnings for a given SED respondent in a given year is computed by adding their yearly W2 and ILBD earnings. We distinguish between university and non-university earnings by linking a list of public-use EINs for all universities, technical schools, and vocational schools in the U.S. to the W2 tax records. This EIN list is available from the Integrated Postsecondary Education Data System (IPEDS), which is maintained by the National Center for Education Statistics (NCES).

2.4 Analysis Sample

Our final analysis sample is comprised of all SED respondents with a unique PIK and non-missing gender information that have positive earnings for all of the first five years after they

[8]The ILBD is the complement of the Longitudinal Business Database (LBD) which is the universe of employer firms in the United States (Chow et al., 2021; Jarmin and Miranda, 2002)

receive their Ph.D. Thus, this sample is composed of individuals who are strongly attached to the US labor market after they receive their degree. The restriction to postive earnings is because our US-based tax data cannot track individuals who leave the United States. This is especially problematic for the highly mobile individuals in the SED. When a respondent is not present in the W2 or 1040-C tax records they may genuinely be out not working and have zero earnings or they may have positive earning outside the US.

2.5 Comparison to SDR

Many of the valuable insights into these questions come from the Survey of Earned Doctorates (SDR), which contains rich, longitudinal data on earnings. We believe that our approach has a number of advantages relative to the SDR for our purposes. First, as (Fealing and Incorvaia, 2022) emphasizes, the SDR is a sample, which makes it difficult to obtain precise estimates for small groups. The population-level data available in the SED matched to tax records addresses this issue. A second concern with the SDR is that earnings are self-reported, which raises issues of non-response bias and misreporting. Here too, tax records are promising for our population, with complete coverage for people, such as the overwhelming majority of PhDs, who are not paid "under the table". Moreover, given that survey non-response tends to be increasing in earnings, there is the potential for our tax records to address non-response bias.

3 Results

In this section, we use our SED-tax linked data to examine earnings disparities between 14 groups defined by seven race categories and two sex categories, which has not heretofore been possible due to small sample sizes from traditional data sources. We begin with gaps by race and sex separately, and then examine the disparities that emerge when race and sex are interacted. We then examine the roles played by field choice and sector of job placement (e.g. industry vs academia) in determining these gaps. Specifically, we add these variables as controls to regressions to see how race and sex gaps are attenuated. We then estimate gaps separately by field and determine where gaps are largest/smallest.

3.1 Gender Gaps

Table XX displays the gender gap in log wages for Ph.D. recipients one year and five years after receiving their degree. The even columns show the raw gaps (conditional on degree year) and the odd columns show the gaps conditional on the following covariates... The odd columns also include Gelbach decompositions showing the contribution of each covariate to the explained log earnings gap.

3.2 Race Gaps

3.3 Ethnicity Gaps

3.4 Interactions: Gender by Race by Ethnicity Gaps

3.5 Differences Across Fields

4 Conclusion

References

Bayer, Amanda and Cecilia Elena Rouse. 2016. "Diversity in the Economics Profession: A New Attack on an Old Problem." *Journal of Economic Perspectives*, 30(4): 221–242.

Ceci, Stephen J., Donna K. Ginther, Shulamit Kahn, and Wendy M. Williams. 2014. "Women in Academic Science: A Changing Landscape." *Psychological Science in the Public Interest*, 15(3): 75–141.

Ceci, Stephen J., Shulamit Kahn, and Wendy M. Williams. 2023. "Exploring Gender Bias in Six Key Domains of Academic Science: An Adversarial Collaboration." *Psychological Science in the Public Interest*, 24(1): 15–73.

Chow, Melissa, Teresa C. Fort, Christopher Goetz, Nathan Goldschlag, James Lawrence, Elisabeth Ruth Perlman, Martha Stinson, and T. Kirk White. 2021. "Redesigning the Longitudinal Business Database." Center for Economic Studies, U.S. Census Bureau Working Papers 21-08.

Fealing, Kaye Husbands and Aubrey DeVeny Incorvaia. 2022. "Understanding Diversity: Overcoming the Small-n Problem." *Harvard Data Science Review*, 4(2).

Foster, Lucia, Erika McEntarfer, and Danielle H. Sandler. 2022. "Diversity and Labor Market Outcomes in the Economic Profession." CES Working Paper No. 22-26.

Ginther, Donna K. and Shulamit Kahn. 2013. "Education and Academic Career Outcomes for Women of Color in Science and Engineering." In *Seeking Solutions: Maximizing American Talent by Advancing Women of Color in Academia*. Vol. , 71–92. National Academies Press.

Ginther, Donna K., Shulamit Kahn, and Walter T. Schaffer. 2016. "Gender, Race/Ethnicity, and National Institutes of Health R01 Awards: Is There Evidence of a Double Bind for Women of Color?" *American Medicine*, 91(8): 1098–1107.

Ginther, Donna K., Walter T. Schaffer, Joshua Schnell, Beth Masimore, Faye Liu, Laurel L. Haak, and Raynard Kington. 2011. "Race, Ethnicity, and NIH Research Awards." *Science*, 333(6045): 1015–1019.

Goetz, Christopher, Zachary Kroff, et al. 2021. "Recent Improvements to the Integrated Longitudinal Business Database (ILBD)." Center for Economic Studies, US Census Bureau.

Jarmin, Ron S. 2007. "Integrated Longitudinal Business Database: Presentation."

Jarmin, Ron S and Javier Miranda. 2002. "The longitudinal business database." *Available at SSRN 2128793*.

Kniffin, Kevin. 2007. "Accessibility to the PhD and Professoriate for First-Generation College Graduates: Review and Implications for Students, Faculty, and Campus Policies." *American Academic*, 3: 49–79.

Li, Diyi and Cory Koedel. 2017. "Representation and Salary Gaps by Race-Ethnicity and Gender at Selective Public Universities." *Educational Researcher*, 46(7): 343–354.

Lundberg, Shelly and Jenna Stearns. 2019. "Women in Economics: Stalled Progress." *Journal of Economics Perspectives*, 33(1): 3–23.

National Center for Science and Engineering Statistics. 2023. "Doctorate Recipients from U.S. Universities: 2022." NSF 24-300, Alexandria, VA: National Science Foundation. Available at https://ncses.nsf.gov/pubs/nsf24300.

Sowell, Robert, Jeff Allum, and Hironao Okahana. 2015. *Doctoral Initiative on Minority Attrition and Completion.* Council of Graudate Schools.

Stansbury, Anna and Robert Schultz. 2023. "The Economics Profession's Socioeconomic Diversity Problem." *Journal of Economic Perspectives*, 37(4): 207–230.

Vilhuber, Lars, Kevin McKinney, et al. 2014. "LEHD Infrastructure files in the Census RDC-Overview." *Center for Economic Studies, US Census Bureau Working Papers*, , (14-26).

Wagner, Deborah, Mary Lane, et al. 2014. "The person identification validation system (PVS): applying the Center for Administrative Records Research and Applications'(CARRA) record linkage software." Center for Economic Studies, US Census Bureau.

Webber, Karen L. ad Canché, Manuel González. 2015. "Not Equal for All: Gender and Race Differences in Salary for Doctoral Degree Recipients." *Research in Higher Education*, 56: 645–672.