Last Place? The Intersection of Ethnicity, Gender, and Race in Biomedical Authorship

By GERALD MARSCHKE, ALLISON NUNEZ, BRUCE A. WEINBERG, AND HUIFENG YU*

In the biomedical sciences, author order reflects the role people play on articles. The first author has primary responsibility for the work, while the last author runs the lab and/or is the principal investigator, who supported the work. Thus, author order affects the credit received for work and conveys information about the stature of authors.

We leverage this feature of scholarly publishing to make two interrelated contributions to our understanding of underrepresentation in the sciences. First, studying the probability that a person is the last author on a publication and algorithmically resolving author ambiguities and imputing ethnicity, gender, and race allows us to use massive, population-level, longitudinal data to study underrepresentation. (West et al. (2013) use a similar approach to study women.)

Second, we use these data to look at ethnicity, gender, race and experience and how they interact in a way that is impossible with sampled data. This analysis is timely because of serious concerns with underrepresentation of women and minorities in biomedicine and other STEM fields (NIH 2012) and with barriers confronting female and minority scientists (e.g. Cook and Kongcharoen 2010, Ginther et al. 2011, and Larivière et al. 2013).

Moreover, research emphasizes the importance of intersectionality, the idea that ethnicity, gender, and race interact to determine experiences and outcomes. For instance, Ong et. al. (2011) identify a "double bind" that particularly disadvantages women from underrepresented racial or ethnic groups. Quantitatively, we distinguish this view from an "additive model" where the difference in outcomes between a non-Hispanic, White man and a woman from an underrepresented racial or ethnic group is given by the sum of coefficients on dummy variables for female

*Marschke: University at Albany, SUNY, 1400 Washington Ave, Albany, NY 12222 (email: marschke@albany.edu). Nunez: University at Albany, SUNY, 1400 Washington Ave, Albany, NY 12222 (email: anunez@albany.edu). Weinberg: 1945 N High St; Columbus, OH 43210 (e-mail: weinberg.27@osu.edu). Yu: University at Albany, SUNY, 1400 Washington Ave, Albany, NY 12222 (email: hyu3@albany.edu). We thank Mikko Packalen for comments and Neil Smalheiser and Vetle Torvik for the use of their data. All authors gratefully acknowledge support from NIA, OBSSR, and NSF SciSIP through P01 AG039347. Weinberg thanks NSF EHR DGE 1348691, 1535399 and the Ewing Marion Kauffman and Alfred P. Sloan Foundations. Weinberg was supported on P01 AG039347 by the NBER directly and on a subaward from NBER to Ohio State. Marschke gratefully acknowledges support from NSF EHR DGE 1661278. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation or the National Institutes of Health. All errors are the responsibility of the authors.

and race and ethnicity. Strikingly, we find that women from some underrepresented groups have better outcomes than those implied by an "additive model," suggesting perhaps a one and a half bind.

II. Data and Methods

A. Data

The core of our data is meta data on 21 million life science articles from 1946 to 2014 in the National Library of Medicine's MEDLINE® 2014 baseline files.

We use the Author-ity data of Torvik and Smalheiser (2009) to measure career age, defined as time since first publication. Author-ity algorithmically identifies roughly 9 million identity clusters (probable people) from the 56,208,832 author-article pairs in MEDLINE through July 2009 with overall recall of 98.8% and precision of 98%.

MEDLINE does not provide author demographic information. We use gender predictions from Genni, developed by Smith, Singh and Torvik (2013). Race and ethnicity are imputed using Ethnicolr, developed by Laohaprapanon and Sood (2017). Ethnicolr uses first and last name to categorize people into four categories that combine race and ethnicity – Hispanic (of any race) and non-Hispanic Asians, Blacks, and Whites.

This piece focuses on U.S. based researchers. To identify author location, we use MapAffil, which provides affiliation information for MEDLINE authors (Torvik 2015). Because location coverage is incomplete, we eliminate all people who are ever outside of the U.S.

Appendix Tables 1 through 3 summarize the variables used in our analysis, data sources, sample deletions, and summary statistics. Our primary sample comprises 1,061,758 author clusters whose careers start after 1947 and last at least 5 years. We focus on all research articles with 2 to 9 authors, leaving 9,266,336 article-author pairs. The mean career age is 11.21. Only 25% of author-article pairs are predicted to be women. For race and ethnicity, the largest group is non-Hispanic White (83%), followed by non-Hispanic Asian (8%), Hispanic (6%) and non-Hispanic Black (3%).

B. Methods

Our main analysis consists of linear regressions of whether author i on a paper j is the last author, $Last_{ij}$:

$$(1)\ Last_{ij} = \beta_0 + \boldsymbol{\beta_1}'\overrightarrow{\boldsymbol{EthGenRace_{\imath}}} + \boldsymbol{\beta_2}'\overrightarrow{\boldsymbol{CareerAge_{\imath\jmath}}} + \boldsymbol{\beta_3}'\overrightarrow{\boldsymbol{X_{\imath\jmath}}} + \varepsilon_{ij},$$

where $\overrightarrow{\boldsymbol{EthGenRace_{\imath}}}$ is a vector of dummy variables giving the ethnicity, gender, and race of author i, $\overrightarrow{\boldsymbol{CareerAge_{\imath\jmath}}}$ is a polynomial in career age. $\overrightarrow{\boldsymbol{X_{\iota\jmath}}}$ is a vector of control variables including a polynomial in publications up to the year before article *j* was published. We also include models where we interact $\overrightarrow{\boldsymbol{EthGenRace_{\iota}}}$ with career age.

III. Results

A. Descriptive Results

Figure 1A shows how author position varies over the career in biomedicine. The probability of being a first author declines from roughly 30% at career ages 0-4 to 16% at career ages 25-29. By contrast, the probability of being a last author increases from 18% to 37%. The probability of being a middle author drops slightly (from 52% to 48%). Thus, while people in our sample are a middle author on roughly half of their pieces, there is a strong pattern of people moving from being first to last author over the career.

Our text focuses on last authorship because it represents the pinnacle of the research career (e.g., Costas and Bordons, 2011). First authorship is mixed in that it indicates primary responsibility for the research, but tends to be subordinate to the last author.

Figure 1B summarizes our most basic findings. The up triangles repeat the last author series from panel A. Blacks (squares) are substantially less likely to be last authors from career ages 5-9 onward, with a gap of 6pp at career ages 25-29. The progression of women (diamonds) and Hispanics (circles) into last authorship is even slower, with a gap of 10pp at career ages 25-29.

B. Regression Analysis

Main Results.— Our basic results in Table 1 show that all groups are substantially less likely to be last authors than non-Hispanic White, men. Column (1) is the most basic specification with controls for career age and its square and year of publication fixed effects. To eliminate differences in papers (e.g. journal quality, article quality, number of coauthors, etc.), column (2) includes article fixed effects. The estimates move closer to zero modestly for women and Hispanics, substantially for Blacks, but become more negative for Asians.

Columns (3) and (4) are analogous but include controls for each author's previous publications and its square. These reduce the estimated gaps relative to the corresponding specifications in columns (1) and (2). The estimates in column (4) show that women are 2.2pp less likely to be last authors and Hispanics are 1.4pp less likely. Column (4) shows that Asians are 2.4pp less likely to be last authors (the estimates without article fixed effects show a 1.3pp gap). The estimates for Blacks are less negative than for the other groups and not statistically significant with article fixed effects.

Appendix Table 4 compares our classification to an alternative source of ethnic classification, the Ethnea model. Appendix Table 5 and Appendix Figure 1 show that these results are robust to imputing ethnicity using the Ethnea model. Additionally, Korean and Japanese authors are moderately less likely to be last author compared to Chinese authors.

Gender Interactions.— Table 2 includes interactions between gender and the race and ethnicity categories. It has the same structure as Table 1. The gender interaction is positive for Hispanics and Blacks, although the significance varies across specifications. Thus, the gender gap is smaller among Blacks and Hispanics than among non-Hispanic Whites. This finding is important because it indicates that the gender gap is not additive with the Black and Hispanic gaps. (F-tests of the joint significance of the interactions between gender and race / ethnicity reported in the table are statistically significant at any conventional level.) As a consequence, while Black and Hispanic women are less likely to be last authors than non-Hispanic White men, the gap is smaller than one would expect given the Black or Hispanic gap and the gender gap, separately. The female interactions for Asians are negative, which is to say that the gender gap among Asians is even larger than the gender gap among non-Hispanic Whites.

The flipside of this less than additive disadvantage for Blacks and Hispanics is that the uninteracted coefficients on Black and Hispanic are more negative in Table 2 than in Table 1, which says that the Black and Hispanic gaps are larger among men than implied by Table 1.

Experience Interactions.— We return here to the life-cycle patterns for each group. The estimates, in Table 3, are organized in the same way as Tables 1 and 2, but we also include estimates with author fixed effects (in columns (3) and (6)). Author fixed effects estimates are valuable for lifecycle analyses because they control for attrition that is related to time invariant differences in productivity (i.e. if the least productive researchers are the most likely to attrit).

The interactions between career age and female are negative, indicating less progression toward last authorship over the career, but estimates with article fixed effects (only) show that women are more likely to be last authors at the very beginning of their careers than others, which was visible in Figure 1B. These estimates are not consistent with the most vulnerable groups (e.g. young women) experiencing the greatest disadvantage, but they do show that women progress toward last authorship more slowly than men.

The interactions between career age and both Hispanic and Black are usually negative (although considerably closer to zero), indicating that Hispanics and Blacks progress more slowly as well. The results for Asians are mixed relative to non-Hispanic Whites. While a positive interaction with career age indicates more rapid progress, insofar as it is associated with a more negative intercept, it also indicates a lower initial level.

Our results are robust to using ethnicity data from Ethnea (Appendix Tables 6 and 7).

IV. Conclusion

Author order is an underutilized way to quantify underrepresentation in the sciences using massive, population-level data. Future work should probe the limits of author order (e.g. if women PIs are more likely to choose not to be listed as last authors), and investigate changes in author order over time. Future work should also investigate the extent to which author order reflects standing in the academic hierarchy, affects promotion, funding, and other outcomes, or both, for researchers generally and also those from underrepresented groups. Future work should also probe the robustness of imputations of gender, race, and ethnicity, especially for women who may change names when they marry.

REFERENCES

Cook, Lisa, and Chaleampong Kongcharoen. 2010. “The Idea Gap in Pink and Black” NBER WP #16331.

Costas, Rodrigo and María Bordons. 2011. “Do age and professional rank influence the order of authorship in scientific publications? Some evidence from a micro-level perspective.” *Scientometrics* 88: 145-161.

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 (No. 6045): 1015-1019.

Laohaprapanon, Surivan and Gaurav Sood, *ethnicolr Algorithm*, https://github.com/appeler/ethnicolr (accessed 9/29/2017)

Larivière, Vincent, Chaoqun Ni, Yves Gingras, Blaise Cronin, and Cassidy R. Sugimoto. 2013. “Bibliometrics: Global gender disparities in science.” *Nature* 504 (No. 7479): 211-213.

NIH. 2012. “Report of the ACD Working Group on Diversity in the Biomedical Research Workforce.”

Ong, Maria, Carol Wright, Lorelle L. Espinosa, and Gary Orfield. 2011. “Inside the Double

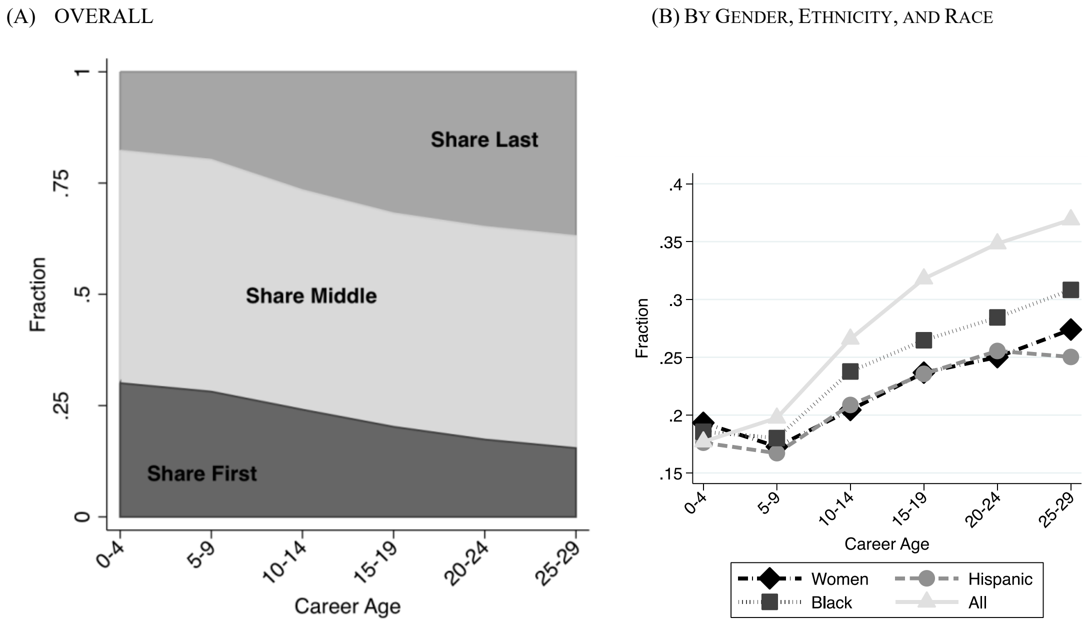
Bind." *Harvard Educational Review* 81 (No. 2, Summer): 172-209.

Smith, Brittany N., Mamta Singh, and Vetle I. Torvik. 2013. "A Search Engine Approach to Estimating Temporal Changes in Gender Orientation of First Names." *JCDL '13 Proceedings of the 13th ACM/IEEE-CS Joint Conference on Digital Libraries*: 199-208.

Torvik, Vetle I.; and Smalheiser, Neil R. 2009. "Author name disambiguation in MEDLINE." *ACM Transactions on Knowledge Discovery from Data*. 3(3): 1-29.

Torvik, Vetle I. 2015. "MapAffil: A bibliographic tool for mapping author affiliation strings to cities and their geocodes worldwide." *D-Lib Magazine*. 21(11-12).

West, Jevin D., Jennifer Jacquet, Molly M. King, Shelley J. Correll, Carl T. Bergstrom. 2013. "The Role of Gender in Scholarly Authorship." *PLoS ONE* 8 (No. 7): e66212



(A) OVERALL

(B) BY GENDER, ETHNICITY, AND RACE

FIGURE 1. AUTHORSHIP BY 5-YEAR CAREER AGE BIN, OVERALL AND BY GENDER, ETHNICITY, AND RACE.

NOTE: The figure is based on researchers who begin publishing between 1980 and 1984, publish for at least five years, and never publish with a non-U.S. Affiliation.

TABLE 1–GENDER, RACE/ ETHNICITY AND BEING LAST AUTHOR

	(1)	(2)	(3)	(4)
Female	-0.0435***	-0.0401***	-0.0340***	-0.0219***
	(0.000772)	(0.000897)	(0.000786)	(0.000948)
Asian	-0.0169***	-0.0310***	-0.0129***	-0.0235***
	(0.00144)	(0.00172)	(0.00136)	(0.00155)
Hispanic	-0.0221***	-0.0205***	-0.0148***	-0.0140***
	(0.00142)	(0.00180)	(0.00143)	(0.00174)
Black	-0.00674***	-0.00164	-0.00486**	-0.000478
	(0.00228)	(0.00250)	(0.00225)	(0.00234)
Career Age and its Square	Y	Y	Y	Y
Year FE	Y		Y	
Article FE		Y		Y
Past Publications and its Square			Y	Y
Observations	9266336	7028707	9266336	7028707
R-squared	0.054	0.252	0.062	0.269

Notes: Observations are author-article pairs. The dependent variable in these least square regressions is defined as 1 if the author is the last author, and as 0 otherwise. Omitted race/ethnic group is White (non-Hispanic). Standard errors (in parentheses) are clustered by article and author.

*** Significant at the 1 percent level.

** Significant at the 5 percent level.

*Significant at the 10 percent level.

TABLE 2 –THE INTERSECTION OF GENDER AND RACE/ ETHNICITY AND BEING LAST AUTHOR

	(1)	(2)	(3)	(4)
Female	-0.0441***	-0.0412***	-0.0337***	-0.0213***
	(0.000865)	(0.000988)	(0.000879)	(0.00104)
Asian	-0.0148***	-0.0309***	-0.00930***	-0.0209***
	(0.00175)	(0.00209)	(0.00165)	(0.00186)
Hispanic	-0.0264***	-0.0246***	-0.0177***	-0.0141***
	(0.00188)	(0.00223)	(0.00190)	(0.00219)

Black	-0.00910***	-0.00412	-0.00662**	-0.00156
	(0.00284)	(0.00311)	(0.00282)	(0.00291)
Female * Asian	-0.00884***	-0.000717	-0.0152***	-0.00982***
	(0.00277)	(0.00310)	(0.00267)	(0.00289)
Female * Hispanic	0.0131***	0.0129***	0.00847***	0.000290
	(0.00271)	(0.00314)	(0.00268)	(0.00302)
Female * Black	0.00916**	0.00932*	0.00682	0.00413
	(0.00439)	(0.00483)	(0.00428)	(0.00459)
F-Stat for Interactions of Female with Asian, Hispanic, and Black	13.62***	6.76***	16.49***	4.32***
Career Age and its Square	Y	Y	Y	Y
Year FE	Y		Y	
Article FE		Y		Y
Past Publications and its Square			Y	Y
Observations	9266336	7028707	9266336	7028707
R-squared	0.054	0.252	0.062	0.269

Notes: Observations are author-article pairs. The dependent variable in these least square regressions is defined as 1 if the author is the last author, and as 0 otherwise. Omitted race/ethnic group is White (non-Hispanic). Standard errors (in parentheses) are clustered by article and author.

*** Significant at the 1 percent level

** Significant at the 5 percent level.

*Significant at the 10 percent level.

TABLE 3 – GENDER, RACE/ETHNICITY AND AUTHORSHIP LIFE-CYCLE PATTERN

	(1)	(2)	(3)	(4)	(5)	(6)
Female	-0.00486***	0.0113***		-0.00692***	0.00762***	
	(0.000896)	(0.00115)		(0.000872)	(0.00110)	
Asian	-0.0142***	-0.0284***		-0.0167***	-0.0287***	
	(0.00157)	(0.00189)		(0.00151)	(0.00178)	
Hispanic	-0.00325**	0.00211		-0.00643***	-0.00714***	
	(0.00159)	(0.00206)		(0.00157)	(0.00199)	
Black	0.00330	0.00322		0.00279	0.00281	
	(0.00244)	(0.00287)		(0.00241)	(0.00281)	
Career Age	0.0165***	0.0249***		0.0123***	0.0170***	
	(0.000112)	(0.000129)		(0.000225)	(0.000311)	
Career Age2	-0.000192***	-0.000315***	-0.000247***	-0.000179***	-0.000300***	-0.000271***
	(0.00000312)	(0.00000343)	(0.00000349)	(0.00000443)	(0.00000552)	(0.00000456)
Career Age * Female	-0.00401***	-0.00516***	-0.00430***	-0.00285***	-0.00301***	-0.00291***
	(0.000110)	(0.000115)	(0.000131)	(0.000103)	(0.000107)	(0.000136)
Career Age * Asian	-0.000148	-0.000138	0.000879***	0.000535***	0.000677***	0.00109***
	(0.000209)	(0.000232)	(0.000260)	(0.000188)	(0.000198)	(0.000257)
Career Age * Hispanic	-0.00186***	-0.00216***	-0.000223	-0.000852***	-0.000668***	0.0000355
	(0.000182)	(0.000202)	(0.000254)	(0.000175)	(0.000193)	(0.000251)
Career Age * Black	-0.000934***	-0.000481*	0.0000464	-0.000722***	-0.000325	-0.000110
	(0.000245)	(0.000275)	(0.000298)	(0.000224)	(0.000245)	(0.000296)
Year FE	Y			Y		
Article FE		Y	Y		Y	Y
Author FE			Y			Y
Past Publications and its Square				Y	Y	Y
Observations	9266336	7028707	6678695	9266336	7028707	6678695
R-squared	0.055	0.254	0.479	0.062	0.269	0.481

Notes: Observations are author-article pairs. The dependent variable in these least square regressions is defined as 1 if the author is the last author, and as 0 otherwise. Omitted race/ethnic group is White. Standard errors (in parentheses) are clustered by article and author.

*** Significant at the 1 percent level

** Significant at the 5 percent level.

*Significant at the 10 percent level.

Last Place: The Intersection between Ethnicity, Gender, and Race in Biomedical Authorship

Appendix

By GERALD MARSCHKE, ALLISON NUNEZ, BRUCE A. WEINBERG, AND HUIFENG YU*

1. Data

1.1 Data and Variables

Our analysis begins with the MEDLINE® 2014 baseline files distributed by the National Library of Medicine (NLM) which contain metadata on over 21 million journal articles (from the most important journals) that publish in the life sciences with a focus on biomedicine, spanning 1946 to 2014.[1] The article metadata in MEDLINE include article title, journal title, publication year, author names, author position, and publication type. We supplement these files with four additional data sources to track authors' careers and identify author race, ethnicity, and affiliation.

Our main outcome variable is whether someone is listed as the last author on a publication. In the biomedical sciences, the first author has primary responsibility for the work, while the last author runs the lab and/or is the principal investigator (e.g. Bhandari et al. 2004; Baerlocher et al. 2007).

1.1.1 Author-ity

We merge into the MEDLINE files the "Author-ity" disambiguation (Torvik, Weeber, Swanson, and Smalheiser 2005; Torvik and Smalheiser 2009) of MEDLINE author names. The resulting dataset presently contains over 9 million identity clusters, that is, (probable) persons, covering MEDLINE records up to July 2009.[2] The Author-ity disambiguation permits the identification of each author's first publication in MEDLINE, and thus the calculation of each author's "MEDLINE career age" or experience.

1.1.2 Race prediction

MEDLINE does not provide demographic information of authors. To impute race we use Ethnicolr, a machine-learning-based classifier trained on a specific data set and implemented in

[1] The most important general science journals such as *Science* and *Nature* that publish life science research are indexed entirely. Others are indexed partially.

[2] The overall recall is 98.8% and precision is about 98%, which while in comparison to other disambiguations at this scale is impressive it means that about 2% of articles belonging to a given investigator are misassigned to a second predicted individual. These splitting errors can occur because of very common names (e.g., John Smith) or radical career changes (an investigator may abruptly change topic areas, affiliations and sets of coauthors). Nonetheless, the Author-ity dataset has already demonstrated broad scientific, social and commercial impact: numerous scholars have obtained the dataset to facilitate their own research, and the National Library of Medicine (NLM) is using the dataset in its PubMed/Entrez/Medline databases as the starting point for a scheme to assign Author IDs to all publications.

Python (Laohaprapanon and Sood 2017). This algorithm assigns persons based on their first and last names to four categories that combine race and ethnicity, specifically Hispanic (regardless of race), non-Hispanic White, non-Hispanic Black, and non-Hispanic Asian. Note that this classification system combines categories that we traditionally think of as representing ethnicity (e.g. Hispanic / non-Hispanic) and race (e.g. Asian, Black, and White). The algorithm was trained on Florida voter registration data from 2017. A name is assigned probabilities of belonging to each of the four classes and among those, the highest probability class is taken as the imputed race.

1.1.3 Genni-Ethnea-Authority

The Genni dataset (Smith, Singh and Torvik 2013) is used to predict the gender of authors covered in the Author-ity data. Genni was trained on the association of names and gender markers generated by Bing.com searches. This dataset contains gender predictions for about 4.6 million authors using first names.[3] Since MEDLINE only supplies full first names for articles published from 2002 onward, the Author-ity data are used to assign first name to records before 2002.

As a robustness check, we compare the results in the text obtained using Ethnicolr to results obtained using Ethnea (Torvik and Agarwal 2016), which infers a name's ethnicity from the frequency of affiliation locations for that name in PubMed using a multinomial logistic model. Ethnea provides a considerably richer classification of ethnicities, employing twenty-six ethnic classes, but can only be used to infer race for people whose ethnicity implies their race (e.g. Chinese names or distinctively Black African names). This dataset identifies the ethnicity of all authors in the Author-ity data.

The size of these data allow us to zoom in on specific groups and look at how ethnicity and gender interact with each other and with experience in a way that simply is not possible with sampled data (Ginther and Kahn 2013).

1.1.4 MapAffil

In this work we focus on authors in the U.S. for two reasons. The first reason for focusing on U.S. authors is that the relationship between ethnicity, gender, race and author order are likely to vary by country. To be concrete, there is no reason to believe that being Black or Hispanic in

[3] They run a logistic regression and use confidence classifications ($p>0.9$ as female, $p<0.1$ as male and unknown otherwise) to increase the accuracy of prediction.

the U.S. is the same as being Black in, for instance, England or Germany and the experiences surely differ compared to being Black in Africa or being Hispanic in a Latin American country or Spain. Similarly, there is little reason to believe that the effects of being a woman are the same across countries.

Our second reason for focusing on U.S. researchers is that MEDLINE indexing outside of the U.S. is less complete and could significantly vary over time. Because we use the first publication to impute career age, it is important that we have thorough coverage. If indexing is more likely to begin mid-career for people working in or moving from a poorly indexed country, we may not accurately measure an author's career age. As an example, it is plausible that the Soviet Union was comparatively closed-off in terms of intellectual innovations, but following the end of the Cold War Russian authors may have migrated to the U.S. where they are indexed in MEDLINE and/or indexing of Russian articles may have improved as tensions eased. Either situation results in these authors entering our sample only following the true beginning of their careers.

To focus on authors from U.S.-based affiliations, we use MapAffil data (Torvik, 2015). This dataset contains predicted affiliation location information of about 31 million article-author pairs from the Author-ity MEDLINE data[4]. We leave authors outside of the U.S. for future work.

1.1.5 Overview

Appendix Table 1 summarizes the main variables used in our analysis along with the data sources.

Our unit of analysis is an article-author pair. Appendix Table 2 summarizes how we arrived at the data set that we use in the analysis. We begin with all article-author pairs covered by Author-ity with valid publication years. We then drop authors with disambiguation errors (e.g. whose career starting and/or ending dates are out of range), and retain only the authors starting their careers between 1947 and 2007. Because Author-ity disambiguates MEDLINE only through July 2009 we exclude articles published after July 2009. MEDLINE provides only the first 10 authors for articles published between 1984 and 1995 and the first 25 authors for articles published between 1996 and 1999. For articles published after 1999, MEDLINE does not truncate author

[4] MapAffil's incorrect location assignments and unresolved ambiguities are rare (< 1%). The incompleteness rate is about 2%, mostly due to a lack of information in the PubMed record's affiliation field.

lists. To address the author truncation problem, we drop from our analysis any article with more than 9 authors. Doing so removes articles produced by very large research teams, for which author order likely has a different meaning than for articles with smaller numbers of authors. Additionally, we only focus on article-author pairs with U.S. affiliations that have valid gender predictions. Thus, we drop authors who ever have a non-U.S. affiliation (other authors on their articles are retained unless they too have ever been outside the U.S.). As a last step, we drop authors with career length less than 5 years. After imposing these restrictions, we are left with 9,266,336 article-author pairs.

1.2 Summary Statistics

Appendix Table 3 presents summary statistics of the variables used in this analysis. In our sample, 24% of all article-author pairs are first authors, 49% are middle authors and 27% are last authors. The mean career age is 11.21 years. Only 25% of author-article pairs are predicted to be women. By Ethnicolr's racial/ethnic classification, the largest group is White (83%), followed by Asian (8%), Hispanic (6%) and Black (3%). According to the Ethnea ethnicity classification, the largest group is English and European (76%), followed by Korean and Japanese (5%), Indian (4%), and Chinese (3%). Within this classification, English and European names tend to have longer careers. Women have shorter careers on average.

Appendix Table 4 reports cross-tabulations of the Ethnicolr ethnicity and race classification and the Ethnea ethnicity classification. The Ethnicolr non-Hispanic Asian category is made up almost entirely (92%) of people Ethnea identifies as Chinese, Indian, Japanese, or Korean. And a substantial majority of people Ethnea identifies as Chinese (75%), Japanese (60%), and Korean (76%) Ethnicolr identifies as non-Hispanic Asian; Indians are split close to evenly between non-Hispanic Asian (46%) and non-Hispanic White (44%). The plurality of names Ethnicolr identifies as Hispanic, Ethnea identifies as Spanish (42%), but 23% of the people Ethnicolr identifies as Hispanic, Ethnea identifies as Italian and 9% Ethnea identifies as French. Close to three quarters of the names that Ethnea identifies as Spanish, Ethnicolr identifies as Hispanic (almost all of the rest are identified as non-Hispanic White). As discussed, Ethnea has little ability to identify Blacks. Fully 61% of the people that Ethnicolr identifies as non-Hispanic Black, Ethnea identifies as having English or French Names; and there is no Ethnea ethnicity that has a high probability of being classified as non-Hispanic Black by Ethnicolr. Among people that Ethnicolr identifies as non-Hispanic White 50% are identified by Ethnea as English, 23% as German, and 9% as French. The vast majority of people identified as Italian, Arabic, English, French, German, and Russian, Ethnicolr identifies as non-Hispanic White. We note that meaningful shares of people Ethnea identifies as Chinese (23%), Indian (44%), Japanese (26%), Korean (19%), and Spanish (25%) are classified by Ethnicolr as non-Hispanic White.

Thus, the three largest inconsistencies between the two classifications are: (1) the lack of a Black category in Ethnea; (2) Ethnicolr identifying as non-Hispanic White meaningful shares of people that Ethnea identifies as Chinese, Indian, Japanese, or Korean; (3) Ethnicolr identifying as Hispanic a meaningful share of people that Ethnea identifies as Italian. At the same time, we view the two classifications as having a moderately high level of consistency.

Appendix Figure 1A shows the trends in last authorship shares over the career using 5-year career age bins based on the Ethnea data for two large ethnic groups (Spanish and non-European), females, and overall. The up triangles repeat the last author series from our main specifications. non-Europeans (squares) are substantially less likely to be last authors from career ages 5-9 onward, with a gap of 8pp at career ages 25-29. The progression of women (diamonds) into last authorship is even smaller, with a gap of 10pp at career ages 25-29. Interestingly, the progression of Spanish (circles) into last authorship, despite being smaller relative to our reference group, is faster than those of non-European ethnicities and women, peaking at career ages 20-24. By career ages 25-29, the gap for Spanish is almost comparable to that of women at about a 9pp gap.

Appendix Figure 1B focuses on three Asian subgroups: Chinese, Indian, and Japanese/Korean. As before, the up triangles represent the trend in our overall sample. Japanese/Koreans (squares) are substantially less likely to be last authors for almost all career ages and relative to all other Asian subgroups, with a gap of about 16pp at career ages 25-29. The progression of Japanese/Korean authors into last authorship is also smaller, with fraction of last authorship rising only about 6pp over the span of 25-29 years. The last authorship shares among both Chinese and Indian authors rise at a more rapid rate than that of Japanese/Korean authors. The progression pattern to last authorship is also very similar for Chinese and Indian authors with both demonstrating a rise in last authorship shares of about 18-19pp while simultaneously being within a 1pp range of each other for each career age bin. Last authorship patterns for Chinese and Indian authors also seem to follow the patterns of the overall sample closely, albeit at lower levels.

2. Analysis using Ethnea

Appendix Tables 5-7 repeat Tables 1-3 using the Ethnea classification of ethnicities. The models are similar to those in the text (see equation (1)), but exclude an explicit race dimension. Chinese, Indian, and Korean or Japanese are not aggregated to explore separate effects within the Asian subgroup. Again, our basic results in Table 1 show that all groups are less likely to be last authors compared to English or European men. Appendix Table 5 shows that these results largely hold using Ethnea data. The most basic specification is Column (1) which includes controls for career age and year of publication fixed effects. Column (2) adds article fixed effects, which for all but the female subgroup makes the coefficient estimates on career age more negative.

Columns (3) and (4) are analogous but include the author's previous publications and its square. Including publications reduces the magnitude of the coefficients relative to the corresponding specifications in columns (1) and (2). The estimates in column (4) show that women are 2.2pp less likely to be last authors and authors with Spanish names 1.2pp less likely. Thus, the results are nearly identical to the results in Table 1 of the main text for women and Hispanics. Of the Asian subgroups, the results in column (4) show that authors with Korean or Japanese names are 3pp less likely to be last authors, compared to 1.5pp for authors with Chinese names. Indians fall in the middle.

The estimates in Appendix Table 6 study interactions between gender and ethnicity and are broadly consistent with the estimates based on Ethnicolr in Table 2. As in the text, we compare our results for gender interactions to those from an "additive model" where the outcomes for women from underrepresented groups are the sum of a dummy variable for women and for the ethnic group. (As in the text, the interactions between gender and ethnicity are statistically significant at any conventional level.) As in the main results, women with Spanish names are more likely to be last authors than one would infer based on the uninteracted gender and Spanish coefficients. Korean and Japanese and Chinese women are less likely to be last authors than implied by an additive model. The results for Indian women are noisy once past publications are included but generally show that Indian women are more likely to be last authors than implied by an additive model. Other ethnicity women are also more likely to be last authors than implied by an additive model.

The estimates in Appendix Table 7 report experience interactions. These estimates are also broadly consistent with the analogous results based on Ethnicolr in Table 3. Women show lower progression to last authorship over their careers than men, as do people with Spanish names, although these results become noisier with the addition of controls. The estimates for Asians in Table 3 show more rapid progression to last authorship, especially once controls are added. Appendix Table 7 shows greater progression for Indians and Chinese (in most specifications), but slower progression for Koreans and Japanese. The estimates for other ethnicity vary by specification.

While the two sets of estimates are not directly comparable, they are broadly reassuring in that they suggest that our main results are not a consequence of the particular approach to imputing ethnicity.

References

Baerlocher, Mark Otto, Marshall Newton, Tina Gautam, George Tomlinson, and Allan S. Detsky. 2007. "The Meaning of Author Order in Medical Research." Journal *of Investigative Medicine* 55(4):175-180.

Bhandari, Mohit, Jason W. Busse, Abhaya V. Kulkarni, P. J. Devereaux, Pamela Leece, and Gordon H. Guyatt. 2004 "Interpreting Authorship Order and Corresponding Authorship." *Epidemiology* 15 (No. 1): 125-126

Ginther, Donna, and Shulamit Kahn. 2013. "Education and Academic Career Pathways for Women of Color in Science and Engineering." In *Seeking Solutions: Maximizing American Talent by Advancing Women of Color in Academia: A Conference Report.* Washington, DC: National Academy Press *4-17.*

Laohaprapanon, Surivan and Gaurav Sood, *ethnicolr Algorithm*, https://github.com/appeler/ethnicolr (accessed 9/29/2017)

Smith, Brittany N., Mamta Singh, and Vetle I. Torvik. 2013. "A Search Engine Approach to Estimating Temporal Changes in Gender Orientation of First Names." *JCDL '13 Proceedings of the 13th ACM/IEEE-CS Joint Conference on Digital Libraries*: 199-208.

Torvik, Vetle I., Marc Weeber, Don R. Swanson, and Neil R. Smalheiser. 2005. "A Probabilistic Similarity Metric for Medline Records: A Model for Author Name Disambiguation." *Journal of the American Society for Information Science and Technology*. 56(2): 140-158.

Torvik, Vetle I. and Neil R. Smalheiser, 2009. "Author name disambiguation in MEDLINE." *ACM Transactions on Knowledge Discovery from Data*. 3(3): 1-29.

Torvik, Vetle I. 2015. "MapAffil: A bibliographic tool for mapping author affiliation strings to cities and their geocodes worldwide." *D-Lib Magazine*. 21(11-12).

Torvik, Vetle I., and Sneha Agarwal. 2016. "Ethnea -- an instance-based ethnicity classifier based on geo-coded author names in a large-scale bibliographic database." International Symposium on Science of Science March 22-23, 2016 - Library of Congress, Washington DC, U.S.A.

APPENDIX TABLE 1: VARIABLE DESCRIPTION

Variables	Description	Data Source
Author Position Indicator		
First	Indicator variable equal to 1 if author is the first author of an article.	Author-ity
Middle	Indicator variable equal to 1 if author is the middle author of an article	Author-ity
Last	Indicator variable equal to 1 if author is the last author of an article.	Author-ity
Demographic Information		
Career Age	Years since he/she published the first article in MEDLINE	Author-ity
Female	Indicator variable equal to 1 if the author's predicted gender is female.	Genni
Asian	Indicator variable equal to 1 if the author's predicted race is Asian.	Ethnicolr (Florida voters)
Hispanic	Indicator variable equal to 1 if the author's predicted ethnicity is Hispanic.	Ethnicolr (Florida voters)
Black	Indicator variable equal to 1 if the author's predicted race is Black.	Ethnicolr (Florida voters)
White	Indicator variable equal to 1 if the author's predicted race is White.	Ethnicolr (Florida voters)
Chinese	Indicator variable equal to 1 if the author's predicted ethnicity is Chinese.	Ethnea
English or European	Indicator variable equal to 1 if the author's predicted ethnicity is English or European.	Ethnea
Indian	Indicator variable equal to 1 if the author's predicted ethnicity is Indian.	Ethnea
Spanish	Indicator variable equal to 1 if the author's predicted ethnicity is Hispanic.	Ethnea
Korean or Japanese	Indicator variable equal to 1 if the author's predicted ethnicity is Korean or Japanese.	Ethnea
Other	Indicator variable equal to 1 if the author's predicted ethnicity is Other.	Ethnea
Other Information		
Past Publications	Accumulated count of all publications through year t-1.	

APPENDIX TABLE 2: SAMPLE SIZE

Sample	Obs.
Authority (with valid publication year)	56,208,832
Disambiguation error[5]	56,195,779
Research article[6]	43,055,616
Multi-author article	40,205,330
Career start between 1947 and 2007[7]	39,354,132
Publication year ≤2009	39,296,245
Team size ≤9[8]	34,638,229
Authors with no non-U.S. affiliation	15,819,319
Has gender prediction	10,939,706
Career length ≥5 years	9,266,336

[5] Negative career age, career end is later than 2009, which is the end year of Author-ity data, or the same author appears more than once in the same paper.

[6] We exclude articles that MEDLINE identifies as “Review”, “English Abstract”, “Case Reports”, “Historical Article”, “Comment”, “Portrait”, “Biography”, “Guideline”, “News” or "Conference”.

[7] We choose 1947 since MEDLINE coverage expands after 1946, although our results are robust to beginning our analysis in 1957. We choose 2007 since Author-ity ends in 2009 and career starts in the data begin to decline in 2008.

[8] In each publication record, MEDLINE lists each author on the publication in order of her appearance and, for some years that we study, truncates the author list at the 10th author.

APPENDIX TABLE 3: SUMMARY STATISTICS

Variables	Mean	Std. Dev.	Source
Observation	9,266,336		
First	0.242	0.428	Author-ity
Middle	0.490	0.500	Author-ity
Last	0.267	0.442	Author-ity
Career Age	11.211	9.721	Author-ity
Female	0.249	0.433	Genni
Asian	0.080	0.272	Ethnicolr
Hispanic	0.062	0.241	Ethnicolr
Black	0.032	0.177	Ethnicolr
White	0.825	0.380	Ethnicolr
Spanish	0.036	0.186	Ethnea
Chinese	0.030	0.172	Ethnea
Indian	0.041	0.198	Ethnea
Korean or Japanese	0.052	0.223	Ethnea
Other	0.065	0.247	Ethnea
English or European	0.775	0.418	Ethnea
Past Publication	24.225	43.651	Author-ity

APPENDIX TABLE 4—RELATIONSHIP BETWEEN ETHNICOLR AND ETHNEA

						Ethnea							
	Chinese	Indian	Japanese	Korean	Spanish	Italian	Arabic	English	French	German	Russian	Other	Total
Ethnicolr													
Non-Hispanic Asian	211,037	173,008	269,841	28,136	1,372	2,668	25,299	9,369	5,183	7,236	6,605	5,192	744,946
Row %	28	23	36	3.78	0	0	3	1	1	1	1	0.7	100
Col %	75	46	60	75.93	0	0	13	0	1	0	2	79.83	8
Cell %	2	2	3	0.3	0	0	0	0	0	0	0	0.06	8
Hispanic (Any Race)	3,630	19,126	37,911	195	243,991	132,794	11,960	19,492	49,251	28,278	28,579	27	575,234
Row %	1	3	7	0.03	42	23	2	3	9	5	5	0	100
Col %	1	5	8.47	0.53	73	23	6	0	6	2	7	0.42	6
Cell %	0	0	0.41	0	3	1	0.13	0	1	0	0	0	6
Non-Hispanic Black	2,848	19,008	24,958	1,822	5,762	8,280	10,784	113,736	70,533	34,602	8,450	28	300,811
Row %	1	6	8.3	0.61	2	3	4	38	23	12	3	0.01	100
Col %	1.01	5.02	5.57	4.92	1.72	1.4	5.61	2.85	9	1.9	2.07	0.43	3.25
Cell %	0	0	0	0	0	0	0	1	1	0	0	0	3
Non-Hispanic White	64,048	167,308	115,011	6,902	83,213	446,141	144,119	3,845,199	658,945	1,748,643	364,559	1,257	7,645,345
Row %	0.84	2.19	1.5	0.09	1.09	5.84	1.88	50.29	8.62	22.87	4.77	0.02	100
Col %	22.75	44.21	25.69	18.63	24.89	75.63	75	96.42	84.06	96.14	89.31	19.33	82.51
Cell %	0.69	1.81	1.24	0.07	0.9	4.81	1.56	41.5	7.11	18.87	3.93	0.01	82.51
Total	281,563	378,450	447,721	37,055	334,338	589,883	192,162	3,987,796	783,912	1,818,759	408,193	6,504	9,266,336
Row %	3.04	4.08	4.83	0.4	3.61	6.37	2.07	43.04	8.46	19.63	4.41	0.07	100
Col %	100	100	100	100	100	100	100	100	100	100	100	100	100
Cell %	3.04	4.08	4.83	0.4	3.61	6.37	2.07	43.04	8.46	19.63	0.07	0.07	100

Notes: Rows percentages shaded in green according to the value relative to the other elements of the row. Column percentages shaded in blue according to the value relative to the other elements of the column. Total percentages shaded in red according to the value relative to other elements of the total.

APPENDIX TABLE 5—GENDER, ETHNICITY AND BEING LAST AUTHOR

	(1)	(2)	(3)	(4)
Female	-0.0441***	-0.0399***	-0.0345***	-0.0217***
	(0.000774)	(0.000898)	(0.000789)	(0.000948)
Spanish	-0.0152***	-0.0183***	-0.00989***	-0.0116***
	(0.00193)	(0.0022)	(0.00191)	(0.00209)
Chinese	-0.0035	-0.0203***	-0.00469**	-0.0149***
	(0.00234)	(0.00236)	(0.00221)	(0.00216)
Indian	-0.00727***	-0.0289***	-0.00585***	-0.0223***
	(0.00194)	(0.00208)	(0.00184)	(0.0019)
Korean or Japanese	-0.0307***	-0.0410***	-0.0225***	-0.0298***
	(0.00174)	(0.00312)	(0.00167)	(0.00278)
Other	-0.00314**	-0.0210***	-0.000452	-0.0195***
	(0.00159)	(0.00197)	(0.00154)	(0.00191)
Career Age and its Square	Y	Y	Y	Y
Year FE	Y		Y	
Article FE		Y		Y
Past Publications and its Square			Y	Y
Observations	9266336	7028707	9266336	7028707
R-squared	0.054	0.252	0.062	0.269

Notes: Observations are author-article pairs. The dependent variable in these least square regressions is defined as 1 if the author is the last author, and as 0 otherwise. Omitted ethnicity group is English or European. Standard errors (in parentheses) are clustered by article and author.

*** Significant at the 1 percent level

** Significant at the 5 percent level.

*Significant at the 10 percent level.

APPENDIX TABLE 6—THE INTERSECTION OF GENDER AND ETHNICITY AND BEING LAST AUTHOR

	(1)	(2)	(3)	(4)
Female	-0.0465***	-0.0412***	-0.0367***	-0.0220***
	(0.000907)	(0.00102)	(0.000915)	(0.00106)
Spanish	-0.0171***	-0.0196***	-0.0108***	-0.0104***
	(0.00255)	(0.00275)	(0.00253)	(0.00262)
Chinese	-0.000576	-0.0200***	-0.000497	-0.0124***
	(0.00304)	(0.00305)	(0.00285)	(0.00273)
Indian	-0.00930***	-0.0316***	-0.00723***	-0.0236***
	(0.00249)	(0.00265)	(0.00235)	(0.00239)
Korean or Japanese	-0.0290***	-0.0413***	-0.0197***	-0.0272***
	(0.00202)	(0.00355)	(0.00194)	(0.00314)
Other	-0.0134***	-0.0239***	-0.0126***	-0.0229***
	(0.00219)	(0.00242)	(0.00211)	(0.00235)
Female * Spanish	0.00625*	0.00438	0.00319	-0.00356
	(0.00365)	(0.00393)	(0.00360)	(0.00374)
Female * Chinese	-0.00964**	-0.000574	-0.0140***	-0.00788*
	(0.00436)	(0.00443)	(0.00418)	(0.00417)
Female * Indian	0.00717*	0.00936**	0.00485	0.00453
	(0.00371)	(0.00398)	(0.00356)	(0.00368)
Female * Korean or Japanese	-0.0105***	0.00107	-0.0167***	-0.00990**
	(0.00360)	(0.00434)	(0.00347)	(0.00409)
Female * Other	0.0313***	0.0115***	0.0371***	0.0134***
	(0.00294)	(0.00365)	(0.00284)	(0.00356)
F-Stat for Interactions of Female with the Ethnicity Variables	27.68***	3.07***	44.25***	5.54***
Career Age and its Square	Y	Y	Y	Y
Year FE	Y		Y	
Article FE		Y		Y
Past Publications and its Square			Y	Y
Observations	9266336	7028707	9266336	7028707
R-squared	0.054	0.252	0.062	0.269

Notes: Observations are author-article pairs. The dependent variable in these least square regressions is defined as 1 if the author is the last author, and as 0 otherwise. Omitted ethnicity group is English or European. Standard errors (in parentheses) are clustered by article and author.

*** Significant at the 1 percent level

** Significant at the 5 percent level.

*Significant at the 10 percent level.

APPENDIX TABLE 7—GENDER, ETHNICITY AND AUTHORSHIP LIFE-CYCLE PATTERN

	(1)	(2)	(3)	(4)	(5)	(6)
Female	-0.00538***	0.0123***		-0.00748***	0.00822***	
	(0.000897)	(0.00115)		(0.000873)	(0.00110)	
Spanish	-0.00655***	-0.00519*		-0.00842***	-0.0101***	
	(0.00230)	(0.00267)		(0.00218)	(0.00253)	
Chinese	-0.0144***	-0.0355***		-0.0145***	-0.0279***	
	(0.00259)	(0.00275)		(0.00252)	(0.00268)	
Indian	-0.0176***	-0.0409***		-0.0191***	-0.0380***	
	(0.00205)	(0.00246)		(0.00200)	(0.00232)	
Korean or Japanese	-0.0163***	-0.0165***		-0.0192***	-0.0227***	
	(0.00195)	(0.00309)		(0.00191)	(0.00297)	
Other	0.0123***	-0.0161***		0.0113***	-0.0189***	
	(0.00169)	(0.00213)		(0.00169)	(0.00225)	
Career Age	0.0165***	0.0249***		0.0122***	0.0169***	
	(0.000113)	(0.000130)		(0.000226)	(0.000311)	
Career Age2	-0.000191***	-0.000314***	-0.000247***	-0.000179***	-0.000300***	-0.000271***
	(0.00000314)	(0.00000344)	(0.00000349)	(0.00000443)	(0.00000552)	(0.00000456)
Career Age * Female	-0.00401***	-0.00522***	-0.00430***	-0.00284***	-0.00305***	-0.00292***
	(0.000111)	(0.000115)	(0.000131)	(0.000104)	(0.000107)	(0.000136)
Career Age * Spanish	-0.000839***	-0.00124***	-0.000222	-0.000156	-0.000150	0.00000621
	(0.000273)	(0.000277)	(0.000308)	(0.000253)	(0.000255)	(0.000305)
Career Age * Chinese	0.00151***	0.00193***	0.000430	0.00130***	0.00156***	0.000571
	(0.000379)	(0.000351)	(0.000366)	(0.000357)	(0.000324)	(0.000363)
Career Age * Indian	0.00112***	0.00127***	0.00177***	0.00137***	0.00157***	0.00186***
	(0.000255)	(0.000271)	(0.000276)	(0.000238)	(0.000246)	(0.000274)
Career Age * Korean or Japanese	-0.00136***	-0.00244***	-0.00123***	-0.000270	-0.000709***	-0.000480

	(0.000223)	(0.000259)	(0.000423)	(0.000209)	(0.000231)	(0.000411)
Career Age * Other	-0.00137***	-0.000421**	0.000802***	-0.00105***	-0.0000642	0.000975***
	(0.000195)	(0.000203)	(0.000230)	(0.000190)	(0.000228)	(0.000239)
Year FE	Y			Y		
Article FE		Y	Y		Y	Y
Author FE			Y			Y
Past Publications and its Square				Y	Y	Y
Observations	9266336	7028707	6678695	9266336	7028707	6678695
R-squared	0.055	0.254	0.479	0.062	0.269	0.481

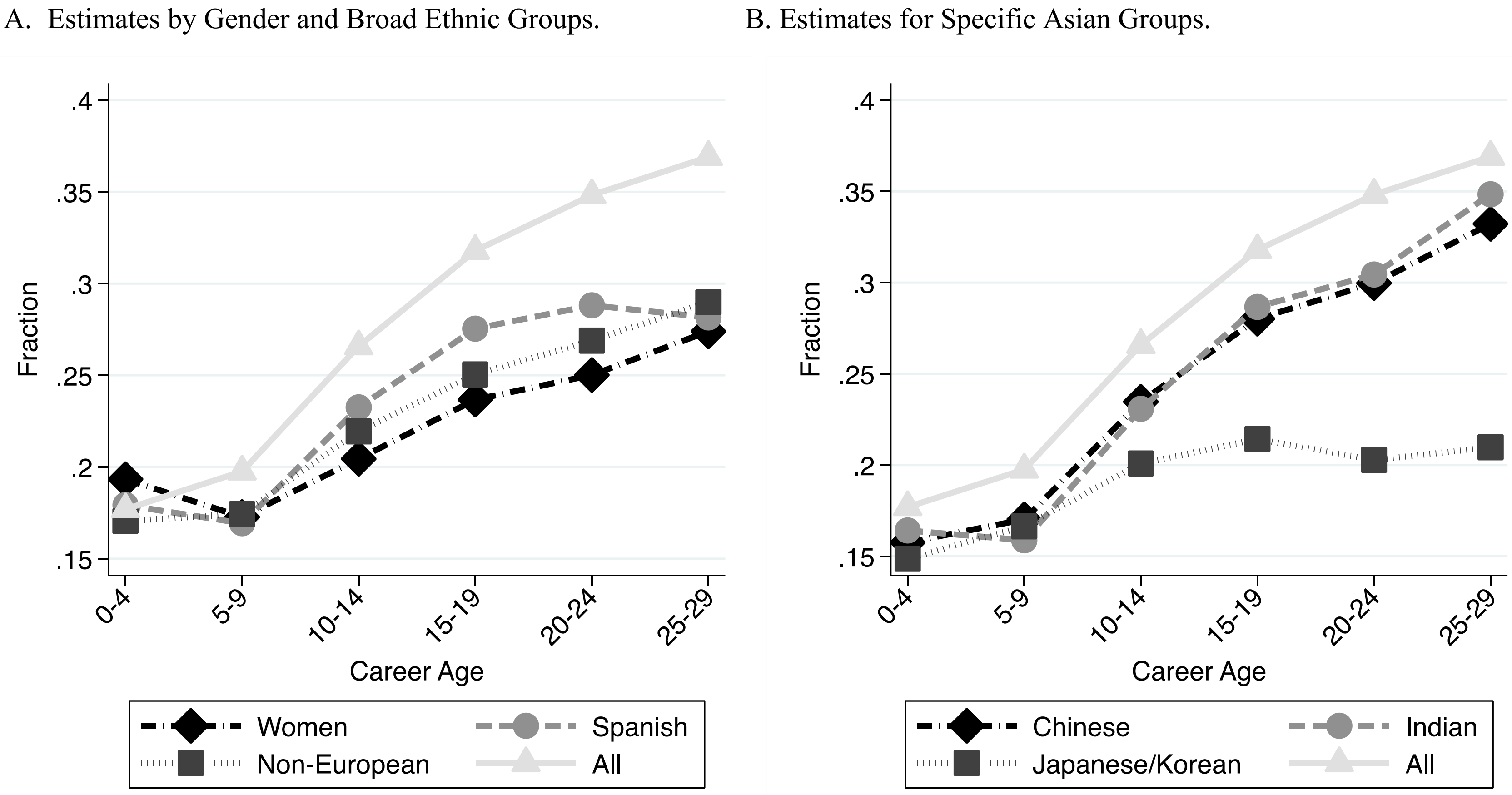
Notes: Observations are author-article pairs. The dependent variable in these least square regressions is defined as 1 if the author is the last author, and as 0 otherwise. Omitted ethnicity group is English or European. Standard errors (in parentheses) are clustered by article and author.

*** Significant at the 1 percent level

** Significant at the 5 percent level.

*Significant at the 10 percent level.

APPENDIX FIGURE 1—AUTHORSHIP BY 5-YEAR CAREER AGE BIN, OVERALL AND BY GENDER AND ETHNICITY


A. Estimates by Gender and Broad Ethnic Groups.

B. Estimates for Specific Asian Groups.

Notes: Estimates from the Ethnea model of ethnicity.