AEA Papers and Proceedings 2018, 108: 222–227
https://doi.org/10.1257/pandp.20181111

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 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.

I. 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 percent and precision of 98 percent.

MEDLINE does not provide author demographic information. We use gender predictions from Genni, developed by Smith, Singh, and

* Marschke: University at Albany, SUNY, 1400 Washington Avenue, Albany, NY 12222 (email: marschke@albany.edu); Nunez: University at Albany, SUNY, 1400 Washington Avenue, Albany, NY 12222 (email: anunez@albany.edu); Weinberg: The Ohio State University, 1945 N High Street; Columbus, OH 43210 (email: weinberg.27@osu.edu); Yu: University at Albany, SUNY, 1400 Washington Avenue, 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.

[†] Go to https://doi.org/10.1257/pandp.20181111 to visit the article page for additional materials and author disclosure statement(s).

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 US 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 United States.

Online 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 percent of author-article pairs are predicted to be women. For race and ethnicity, the largest group is non-Hispanic white (83 percent), followed by non-Hispanic Asian (8 percent), Hispanic (6 percent), and non-Hispanic black (3 percent).

B. *Methods*

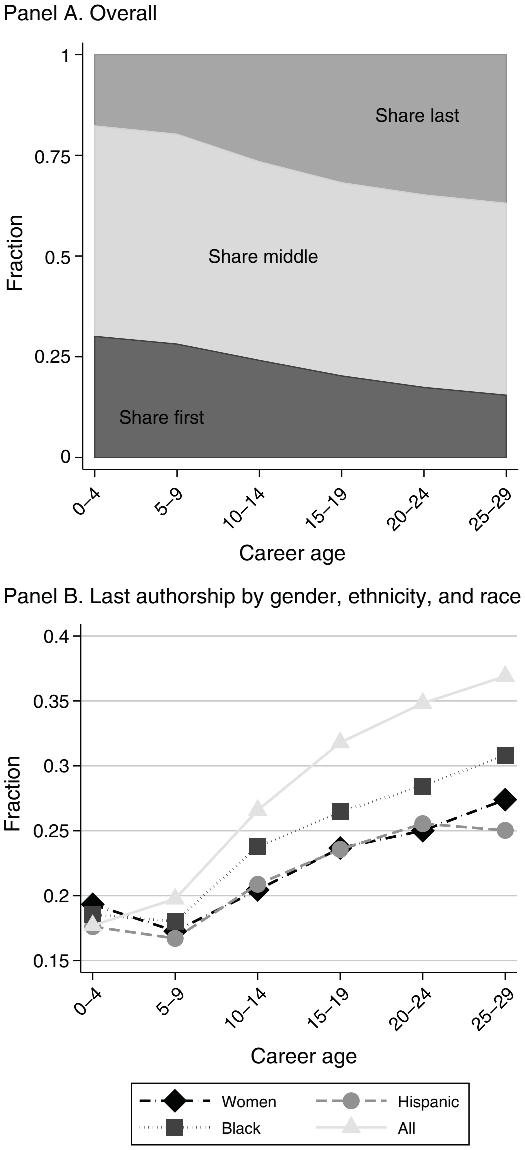
Our main analysis consists of linear regressions of whether author *i* on a paper *j* is the last author, $Last_{ij}$:

$$Last_{ij} = \beta_0 + \boldsymbol{\beta}_1' \overline{\mathbf{EthGenRace}_i} + \boldsymbol{\beta}_2' \overline{\mathbf{CareerAge}_{ij}} + \boldsymbol{\beta}_3' \overline{\mathbf{X}_{ij}} + \varepsilon_{ij}, \tag{1}$$

where $\overline{\mathbf{EthGenRace}_i}$ is a vector of dummy variables giving the ethnicity, gender, and race of author *i*, $\overline{\mathbf{CareerAge}_{ij}}$ is a polynomial in career age. The variable $\overline{\mathbf{X}_{ij}}$ 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 $\overline{\mathbf{EthGenRace}_i}$ with career age.

II. Results

A. *Descriptive Results*


Panel A. Overall

Panel B. Last authorship by gender, ethnicity, and race

Figure 1. Authorship by Five-Year Career Age Bin: Overall and by Gender, Ethnicity, and Race

Notes: 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-US Affiliation. Panel A shows the fraction of author-article pairs by first, middle, and last authorships and age. Panel B shows the fraction of last authorships by demographic group.

Figure 1, panel A, shows how author position varies over the career in biomedicine. The probability of being a first author declines from roughly 30 percent at career ages 0–4 to 16 percent at career ages 25–29. By contrast, the probability of being a last author increases

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	Yes	Yes	Yes	Yes
Year fixed effects	Yes		Yes	
Article fixed effects		Yes		Yes
Past publications and its square			Yes	Yes
Observations	9,266,336	7,028,707	9,266,336	7,028,707
R^2	0.054	0.252	0.062	0.269

Notes: Observations are author-article pairs. The dependent variable in these least squares 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.

from 18 percent to 37 percent. The probability of being a middle author drops slightly (from 52 percent to 48 percent). 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 1, panel B, 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 6 percentage points at career ages 25–29. The progression of women (diamonds) and Hispanics (circles) into last authorship is even slower, with a gap of 10 percentage points 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 2.2 percentage points less likely to be last authors and Hispanics are 1.4 percentage points less likely. Column 4 shows that Asians are 2.4 percentage points less likely to be last authors (the estimates without article fixed effects show a 1.3 percentage point gap). The estimates for blacks are less negative than for the other groups and not statistically significant with article fixed effects.

Online Appendix Table 4 compares our classification to an alternative source of ethnic classification, the Ethnea model. Online Appendix Table 5 and online 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

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

	(1)	(2)	(3)	(4)
Female	−0.0441 (0.000865)	−0.0412 (0.000988)	−0.0337 (0.000879)	−0.0213 (0.00104)
Asian	−0.0148 (0.00175)	−0.0309 (0.00209)	−0.00930 (0.00165)	−0.0209 (0.00186)
Hispanic	−0.0264 (0.00188)	−0.0246 (0.00223)	−0.0177 (0.00190)	−0.0141 (0.00219)
Black	−0.00910 (0.00284)	−0.00412 (0.00311)	−0.00662 (0.00282)	−0.00156 (0.00291)
Female × Asian	−0.00884 (0.00277)	−0.000717 (0.00310)	−0.0152 (0.00267)	−0.00982 (0.00289)
Female × Hispanic	0.0131 (0.00271)	0.0129 (0.00314)	0.00847 (0.00268)	0.000290 (0.00302)
Female × black	0.00916 (0.00439)	0.00932 (0.00483)	0.00682 (0.00428)	0.00413 (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	Yes	Yes	Yes	Yes
Year fixed effects	Yes		Yes	
Article fixed effects		Yes		Yes
Past publications and its square			Yes	Yes
Observations	9,266,336	7,028,707	9,266,336	7,028,707
R^2	0.054	0.252	0.062	0.269

Notes: Observations are author-article pairs. The dependent variable in these least squares 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.

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 flip side 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 life-cycle 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 article fixed effects) show that women are more likely to be last authors at the very beginning of their careers than others, which was visible in

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 age^2	−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 fixed effects	Yes			Yes		
Article fixed effects		Yes	Yes		Yes	Yes
Author fixed effects			Yes			Yes
Past publications and its square				Yes	Yes	Yes
Observations	9,266,336	7,028,707	6,678,695	9,266,336	7,028,707	6,6786,95
R^2	0.055	0.254	0.479	0.062	0.269	0.481

Notes: Observations are author-article pairs. The dependent variable in these least squares 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.

Figure 1, panel B. 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 (online Appendix Tables 6 and 7).

III. 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." National Bureau of Economic Research Working Paper 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 (1): 145–61.

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): 1015–19.

Laohaprapanon, Suriyan, and Gaurav Sood. 2017. "Impute Race and Ethnicity Based on Name." GitHub. https://github.com/appeler/ethnicolr (accessed September 29, 2017).

Larivière, Vincent, Chaoqun Ni, Yves Gingras, Blaise Cronin, and Cassidy R. Sugimoto. 2013. "Bibliometrics: Global Gender Disparities in Science." *Nature* 504 (7479): 211–13.

National Institutes of Health. 2012. *Biomedical Research Workforce Working Group Report: A Working Group of the Advisory Committee to the Director*. Bethesda, MD: National Institutes of Health.

Ong, Maria, Carol Wright, Lorelle L. Espinosa, and Gary Orfield. 2011. "Inside the Double Bind: A Synthesis of Empirical Research on Undergraduate and Graduate Women of Color in Science, Technology, Engineering, and Mathematics." *Harvard Educational Review* 81 (2): 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." In *JCDL 2013—Proceedings of the 13th ACM/IEEE-CS Joint Conference on Digital Libraries*, 199–208. New York: Association for Computing Machinery.

Torvik, Vetle I. 2015. "MapAffil: A Bibliographic Tool for Mapping Author Affiliation Strings to Cities and Their Geocodes Worldwide." *D-Lib Magazine: The Magazine of the Digital Library Forum* 21 (11/12).

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

West, Jevin D., Jennifer Jacquet, Molly M. King, Shelley J. Correll, and Carl T. Bergstrom. 2013. "The Role of Gender in Scholarly Authorship." *PLoS ONE* 8 (7): e66212.