Journal of Informetrics 11 (2017) 835–837

Contents lists available at ScienceDirect

Journal of Informetrics

journal homepage: www.elsevier.com/locate/joi

Correspondence

Comment on "Does your surname affect the citability of your publications"☆

1. Introduction

The question whether the position of a researcher's last name in the alphabet matters for his or her scientific career is important. Abramo and D'Angelo (2017) attempt to answer this question and claim that they find "no evidence whatsoever that alphabetically earlier surnames gain advantage". This stands in contrast to almost all literature on the topic (for a recent survey, see Weber, 2016). The reason that the results in Abramo and D'Angelo (2017) differ from the standard results is that the way they conduct their research is problematic and their claims are unsubstantiated.

2. Problems with mere correlations

One of the main problems is that Abramo and D'Angelo (2017) look at mere correlations. One cannot just look at the correlation of some measure of success (e.g. citations, as used in their article) and last name initial and assume that it will be high if alphabetical discrimination exists without keeping other things equal. Naturally, if there is a very heterogeneous population (consisting of scholars with PhDs from MIT and other top schools, others with PhDs from Italian country-side universities, some older scholars, some younger ones, some at top places with great co-authors, some working at places with hardly any research activity, etc.), the first letter of the last name will not be a main driving factor of citations and the correlations will at best be extremely low! Therefore just looking at correlations is not useful. Note that even if the theoretical distributions of the mentioned explanatory variables were absolutely identical across last name initials (which is a best-case scenario), the correlation of citations and last name initials would still be very low, because there is a lot of variation within each last name initial.[1]

Thus, considering only correlations can be problematic even in the best-case scenario in which no systematic biases are present. However, systematic biases, such as an omitted-variable bias, could additionally corrupt the results. An omitted-variable bias exists if an explanatory variable is neglected that is correlated with both the outcome variable and with the explanatory variable under consideration. In the study by Abramo and D'Angelo (2017), such an omitted variable could for example be ethnicity. It is not unreasonable to think that Asian scholars in Italy perform better than Italian scholars (why would Italian universities otherwise hire them given that there may be troubles with missing Italian language skills) while their last names come on average later in the alphabet. It could well be the case that alphabetical discrimination exists, but that it is not detected by the correlations or that the correlations point toward the opposite, because Asian scholars in Italy perform better than other scholars. To be fair, the authors have removed some surname initials for which there is probably a very uneven distribution of ethnicities. However, they have not removed all. P and S are for example still in the sample and

☆ Disclaimer: The views expressed are those of the author and do not necessarily reflect those of the Bank of Lithuania.

[1] As an illustration of the last point, consider the following simplistic example. Two (sub)samples, S1 and S2 are drawn from normal distributions with the same standard deviation. The distributions are identical except for a shift in the mean (where the size of the shift in the mean is low compared to the standard deviation of the distributions). To be concrete, consider 1000 observations both in S1 and S2, a mean of the normal distribution of 0 in S1 and 0.1 in S2 and standard deviations of 10. Such a scenario can easily be simulated and the correlation between a binary indicator (0 for S1 and 1 for S2) and the drawn values in S1 and S2 can be calculated. The results from 10000 simulations are as follows. The Spearman rank-correlation is on average (across all simulations) less than 0.005 and always between −0.086 and 0.089. It is negative slightly more than 40 percent of the times. The correlations are statistically significantly different from zero less than 6 percent of the times (at the conventional 5-% level); in more than a quarter of the cases in which the test rejects the hypothesis of zero correlation the correlation is negative. Thus, even in this best-case scenario in which the theoretical distributions are exact shifts of one another and with a relatively large number of observations, the correlation is generally close to zero and statistically mainly non-significant. One can easily verify that also in less simplistic settings mere correlations are usually not a useful measure if uncontrolled factors other than the investigated explanatory variable lead to a large variation in the data (one could for example consider discrete and non-negative distributions or build a full model in which a set of explanatory variables explains the number of citations together with last name initials).

http://dx.doi.org/10.1016/j.joi.2017.06.002

Patel and Singh are the most common surnames in India with together more than 100 million people carrying these names (according to Wikipedia). Removing all letters with an uneven name distribution is of course neither possible nor desirable. I also do not want to claim that Asians in the sample are driving the effects (maybe there are only very few of them in the sample); this is just one example of something that might have gone wrong. The point is to illustrate that mere correlations are a very simplistic tool not capturing many possible effects (a regression analysis controlling for variables such as ethnicity would have been a more reasonable approach).

It is also problematic to look at citations per scholar, because the distribution of scholars in academia may be endogenous. Weak scholars with last names late in the alphabet might for example be more likely to leave academia than weak scholars with last names early in the alphabet. In such a case correlations between citations and last names would also underestimate alphabetical discrimination (this problem would persist when conducting a simple regression analysis).

3. Problems with the sample

The sample analyzed by the authors has some weaknesses. The authors attempt to examine alphabetical discrimination due to alphabetical name ordering and therefore exclude the life sciences (they write that they "consider all of the publications by all professors in sciences of Italian universities, with the exception of the life sciences where standard practice in Italy is to order the byline according to the authors' intellectual contributions"). However, the authors do not examine a single field by itself in which alphabetical name ordering is the norm (there are only two full disciplines that predominantly adhere to an alphabetical name ordering norm, namely economics and mathematics; see Waltman, 2012).

The authors exclude economics from their analysis. Economics is a particularly interesting case, because in economics most forces are at work that can lead to alphabetical discrimination. The use of the 'et al.' notation is very common in economics and less common in mathematics. Reference lists are in general ordered alphabetically in economics while they are often ordered according to first appearance in the text in mathematics. That economics is not part of the analysis is not the only problem with the data: mathematics is grouped together with computer science, which is a field in which the contribution-based rule is common.

Thus, the authors base their findings to a great extent on various fields in which mainly a contribution-based ordering norm exists. Of course, also in these fields, some alphabetical ordering exists, but some more alphabetical ordering appearing than what a clean contribution-based rule would suggest is something completely different from having an actual alphabetical name ordering norm (the 'excess' alphabetical ordering appearing in fields with a contribution-based norm could for example be due to tie-breaking when all authors contributed equally, to interdisciplinary collaborations with scholars working in fields where an alphabetical norm prevails, or to – academically probably not very strong – researchers not even knowing the author ordering norm in their field). Defendants of the study might claim that author ordering norms in Italy are different from other parts of the world and that more scholars in Italy make use of an alphabetical norm. However, this is true at most to a limited extent. Abramo and D'Angelo (2017) report the percentage of publications in which the first author is the one coming first in the alphabet (it seems strange that they report numbers for the first author coming first in the alphabet instead of numbers of articles with a fully alphabetical ordering, which are necessarily lower; an article with the author order 'A-name, Z-name, M-name' cannot be considered to adhere to an alphabetical norm). Given that the numbers they report are overstatements of alphabetical name ordering and seeing that these overstatements are still below one half for the majority of categories they consider (and only much above one half for one category), the described problem of the sample persists to a large extent.

4. Concluding remarks

For all the reasons described above, the analysis by Abramo and D'Angelo (2017) does not tell us much. What is interesting, however, is that in the merged category mathematics and computer science the correlation coefficient increases from 0.004 to 0.057 when looking only at the top 10 percent of researchers. 10 percent of all Italian scholars in a field are still a lot of researchers. Given that alphabetical discrimination without additional controls can be better detected at the very top (where the quality of scholars is more homogeneous than in the whole sample and where a pass-through effect of the best scholars with early last name initials moving to a better 'category' is impossible) and given that mere correlations are not suitable to detect such discrimination, this tenfold increase in correlation when looking at the top 10 percent points toward the existence of discrimination rather than anything else (one should be extremely careful when interpreting these numbers, though).

In summary, in contrast to what the authors claim, the article by Abramo and D'Angelo (2017) does not present any analysis supporting the claim that alphabetical discrimination does not exist or that it is overrated.

References

Abramo, G., & D'Angelo, C. A. (2017). Does your surname affect the citability of your publications? *Journal of Informetrics*, *11*(1), 121–127.
Waltman, L. (2012). An empirical analysis of the use of alphabetical authorship in scientific publishing. *Journal of Informetrics*, *6*(4), 700–711.
Weber, M. (2016). *The effects of listing authors in alphabetical order: A survey of the empirical evidence*. Bank of Lithuania Occasional Paper no 12/2016.

Matthias Weber*
CEFER, Bank of Lithuania & Faculty of Economics, Vilnius University, Lithuania

* Correspondence to: Research Center CEFER (Bank of Lithuania), Totoriu g. 4, Vilnius, 01121, Lithuania.
E-mail addresses: mweber@lb.lt, matthias.weber@ef.vu

8 June 2017
Available online 29 June 2017