Citations, journal ranking and multiple authorships reconsidered: evidence from almost one million articles

Wohlrabe, Klaus and Sommer, Vera

August 2016

Citations, journal ranking and multiple authorships reconsidered: evidence from almost one million articles

Abstract

In this paper we reconsider the investigation by Moosa (2016) using a much larger data set of almost one million articles listed in RePEc. This article provides new insights into the effects of co-authorship on citation counts and the correlation between quality of papers and quality of the publishing journal. Our evidence is partially in contrast to the results reported in Moosa (2016). We find a positive correlation between the h-index of a journal and the quality of papers measured in terms of citations. This correlation becomes almost perfect using a non-linear model. Results from a regression of citation counts on the number of authors show evidence of a positive and significant effect of co-authorship on the quality of a paper when time effects and large sets of top-cited articles are taken into account. The inclusion of time effects and the large data set, that allows to differentiate between top-cited cohorts, add further insights to the existing literature.

Vera Sommer
Ifo Institute – Leibniz-Institute
for Economic Research
at the University of Munich e.V.
Poschingerstr. 5
81679 Munich, Germany
Phone: +49(0)89/9224-1229

Klaus Wohlrabe*
Ifo Institute – Leibniz-Institute
for Economic Research
at the University of Munich e.V.
Poschingerstr. 5
81679 Munich, Germany
Phone: +49(0)89/9224-1229
wohlrabe@ifo.de

* Corresponding author.

JEL Code: A12, A14
Keywords: Citations, multiple authorship, journal quality, RePEc

1 Introduction

Given the growing importance of citation counts for economists, two questions are frequently raised: Is there an effect of a paper's number of co-authors on its citations? Are top papers necessarily published in top journals? In the recent past, several researchers have analyzed these questions. While there is wide agreement on the fact that there is a tendency towards multi-authored papers over the last decades, see for instance Sutter and Kocher (2004), Nowell and Grijalva (2011) or Rath and Wohlrabe (2016a), the effect of this phenomenon on the number of citations are controversially discussed. Using information about 300 top papers in economics from RePEC, Moosa (2016) analyzes whether first, the link between the quality of a journal and the quality of the papers that are published therein and second, a positive relationship between the number of authors and the number of citations are confirmed by the data. He concludes that there is a correlation between top-cited articles and top journals, which is, however, not perfect. Moreover, based on his data set he does not find a significant link between co-authorship and citations. Therefore, he deduces that multi-authored papers are not (necessarily) better in quality. In this paper we challenge the results of Moosa (2016) using a much larger data set comprising almost one million articles from RePEc. There are three critical points in the analysis of Moosa (2016). First, the choice of the top 300 cited papers is ad hoc. It might be the case that the conclusions change when one considers the top 500 or top 1,000 papers. Second, Moosa (2016) does not account for the time dimension in explaining citations. It is common sense that older papers have much more time to gather citations than more recently published articles. A related issue is that many top-cited papers were written in times when solo-authorship was quite common. We show in this paper that there is a significantly positive relationship between citations and the number of authors. Third, we use a consistent data set. Moosa (2016) uses citation data from RePEc, whereas the journal quality is measured by the SCImago Journal Rank (SJR) and the h-index. Both indicators were obtained using citation data from Scopus. As the citation coverage, both in quality and quantity, for these two databases differ the results of Moosa (2016) might be potentially biased. In our analysis we use only data obtained from the RePEc website.

The article is organized as follows: first, we introduce our data set from RePEc and give some descriptive statistics. Then, the correlation between top papers and top journals is examined. In section 4 we analyze the effect of co-authorship on citations by testing mean differences and regressing citations on the number of co-authors with and without controlling for time-effects. Finally, we conclude by contrasting our main results with Moosa (2016)'s findings.

2 Data

We extracted our data from RePEc (Research Papers in Economics, www.repec.org). In economics, RePEc has become an essential source for the spread of knowledge and ranking of individual authors and academic institutions. RePEc is based on the "active participation principle", i.e. that authors, institutions and publishers have to register and to provide information to the network. This approach has the main advantage that a clear assignment of works and citations to authors and articles is possible. Indeed, the RePEc story has become a success, with more than 45,000 registered authors with listed works and 2,500 journals in economic sciences worldwide as of August 2016. Using a unique identifier, we downloaded all meta-information for more than 1,000,000 journal articles listed in RePEc. This includes the title, the journal, number of authors and citations. Additionally, we restricted ourselves to data up to 2013, as for 2014 not all information for all journals were available. We also excluded articles with obviously misclassified bibliometric information. All data were downloaded on 01/19/2015.[1] Finally, we have data for 953,266 journal articles published in 1895 journals.

The quality level of a journal is captured by the simple impact factor as well as by the h-index. Based on the data set we calculated both indices for all journals. The definition is similar to the "official" impact factor published by Thomson Reuters Journal Citation Reports. The main difference is the year and article coverage of citation counts. In RePEc, all citations

[1] This is the same database which is also been used in Rath and Wohlrabe (2016a) and Rath and Wohlrabe (2016b).

are related to total number of registered articles in a journal. For further details on RePEc see Zimmermann (2013) and Seiler and Wohlrabe (2012).

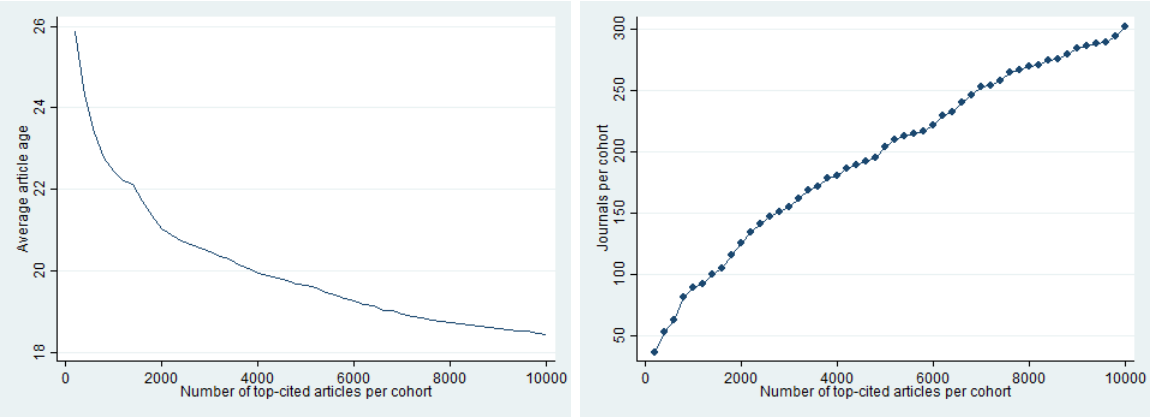
In Table 1 we list the top 20 journals in terms of overall citations. Moreover, the total number of citations, the citations of the top-cited article per journal, the impact factor (IF), the h-index, and, finally, the number of top-cited papers among the top-cited paper cohorts are given. In this ranking the *American Economic Review* is the first journal in terms of overall citations with more than 305,000 citations in total. While its impact factor is relatively low and reflects the high number of papers published, it has the highest h-index of the top 20 journals. The top-cited article in our data set is the paper written by Arellano and Bond (1991) in the *Review of Economic Studies* which has been cited 4,548 times as of January 2015. Up to the top 8 journals, our ranking includes the same journals as Moosa (2016). The ordering of the journals, however, differs significantly. The differences become even more evident when more journals up to the top 20 are included. For instance, the journal *Experimental Economics* is ranked on position 19 in the former but only on position 102 in our ranking. These differences are driven by the much larger data set considered.

Figure 1 plots the average article age as well as the number of journals per cohort against the size of the cohort, where cohort stands for the respective papers that belong to the 200 top-cited articles, the 400 top-cited articles, ..., the 10,000 top-cited articles. We define these cohorts in order to investigate whether the results are driven by the size of a cohort. The left hand panel shows that the more top papers are included, the lower is the average article age. This negative relationship is reasonable as a recently published paper cannot be cited as often as an older one simply because of the lack of time. This finding is further evidence to account for the time dimension when analyzing the effect of multiple authorship on citations. The right hand panel of Figure 1 illustrates that a larger set of top-cited articles coincides with an increase in the number of publishing journals. This is in line with the results reported in Oswald (2007) who documents that top-cited articles are also published in lower ranked journals. Seiler and Wohlrabe (2014) show that for almost all journals in economics the citation distribution is skewed, i.e. dominated by the respective top-cited article.

Table 1: Desciptive statistics for the Top 20 journals in terms of citation count

						Number of top-cited papers				
Journal	Papers	Citations	Top Cited Paper	IF	h-index	200	400	600	800	1000
American Economic Review	8811	305201	2336	34.64	233	20	42	68	92	117
Econometrica	3362	221597	3793	65.91	208	39	59	76	94	111
Journal of Political Economy	3674	208500	3421	56.75	216	35	62	82	100	125
The Quarterly Journal of Economics	2193	167606	3111	76.43	206	20	42	70	82	95
Journal of Finance	5585	141372	1572	25.31	174	12	21	37	49	56
Journal of Econometrics	3541	112390	3519	31.74	137	10	16	28	38	43
Journal of Financial Economics	2200	97280	3170	44.22	139	6	17	24	33	44
Journal of Monetary Economics	2313	93862	4076	40.58	136	11	17	23	30	38
Review of Economic Studies	2262	89626	4548	39.62	133	7	14	20	28	34
Economic Journal	3349	78373	954	23.40	117	3	7	9	12	13
The Review of Economics and Statistics	3753	77735	641	20.71	116	0	2	6	13	19
Journal of Public Economics	3267	74054	808	22.67	105	1	4	5	6	10
Journal of Economic Theory	3757	73716	927	19.62	105	2	8	9	13	14
Journal of International Economics	2629	62000	1058	23.58	112	2	7	9	12	19
European Economic Review	3359	59760	881	17.79	103	1	1	5	5	6
Journal of Economic Perspectives	1479	58831	1028	39.78	127	1	3	7	10	16
Journal of Economic Literature	798	58454	1781	73.25	130	5	17	26	35	45
Journal of Development Economics	2754	51229	424	18.60	100	0	0	0	2	3
Economics Letters	8849	48591	820	5.49	65	1	2	4	4	5
Journal of Banking & Finance	3884	43348	393	11.16	79	0	0	0	1	2

Figure 1: Number of journals and average article age per cohort



3 Top-cited articles vs. journal quality

In order to assess the correlation between the quality of a paper and the quality of the publishing journal, Figure 2 shows the scatterplots between the top-cited papers per journal, the journal impact factor and the h-index. The two latter quantities serve as a measure for the quality of the journal. It is obvious that there is a positive relationship between all three quantities. The correlation between the measures is always larger than 0.79, i.e. high-quality journals tend to have more top cited papers. The lowest interdependence is between the impact factor and the top-cited articles per journal. The correlation between the h-index and the top-cited articles is above 0.80. After studying the scatterplot in the upper right corner, one can see that the relationship between the h-index and the top-cited articles per journal is probably non-linear.

In a next step we follow Moosa (2016) and run a regression of the top-cited articles on the impact factor and the h-index. In Figure 3 we plot the R^2 as a measure of goodness of fit for these regressions for different top-cited cohorts. Additionally, we also allow for a non-linear relationship by adding a squared term of the quality measure. For both quality measures the non-linear model is able to explain larger shares of the total variation in the quality of the paper in terms of citations. Whereas the linear and non-linear models with the impact factor

as explanatory variable lead to an R^2 below 60%, the h-index is able to explain more of the papers' quality. Between the top 200 and top 600 cohorts there is a sharp increase in the R^2, i.e the correlation between quality of papers and journal quality becomes stronger with larger cohorts. A possible reason might be that in a smaller set of top-cited articles there are more outliers, i.e. top-cited articles that are published in journals that have relatively low h-indices. In comparison to Moosa (2016), our larger data set gives us the advantage to differentiate between top-cited groups. Even though the lowest correlation between the h-index and the quality of papers can be found for the top 200 cohort, it is larger than Moosa (2016) already to start with. Moreover, taking a non-linear model and extending the number of included top articles to 2,000 lead to an R^2 of almost one and hence, in contrast to Moosa (2016)'s statement, to correlation close to perfect.

Figure 2: Relationship between top-cited paper, impact factor and h-index

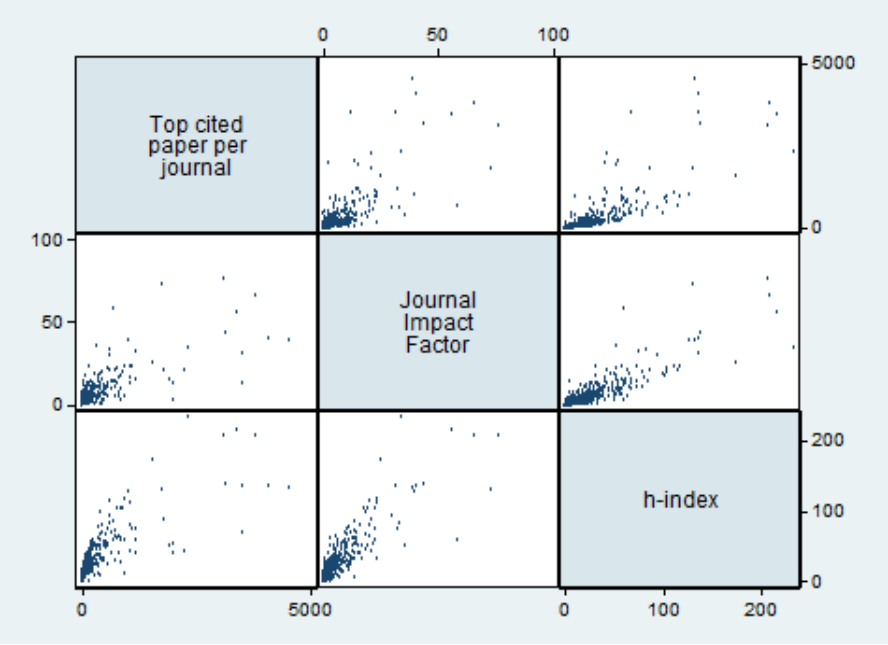
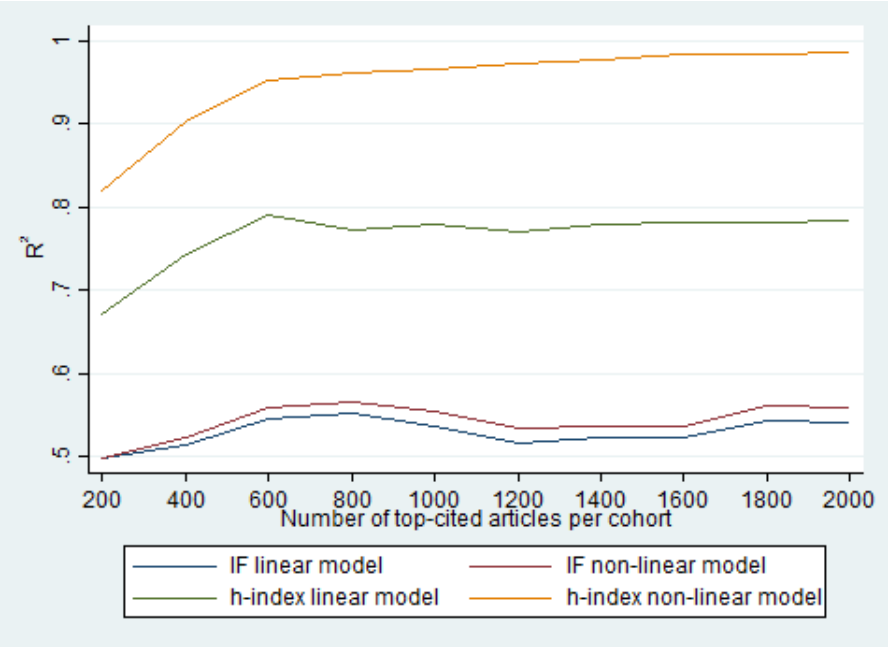


Figure 3: Relationship between cohort and quality indices for journals: Goodness of fit



4 Citations vs. authorship

The number of citations and the number of authors are analyzed next. In the left panel of Figure 4 we plot the distribution of authors and citations per article for the full sample. At first sight, it seems as if more than four authors are in general disadvantageous for the number of citations per article. However, as the right panel of Figure 4 illustrates, the correlation between citations and authorship has increased over the past 60 years. It has been positive since the 1970s, even though the correlation coefficient in absolute terms has always remained small.

Following Moosa (2016), we performed pairwise tests for the mean difference of citations for single-authorship vs. two, three, and four authors as well as two vs. three and four authors, and finally, three vs. four authors, respectively. Again, we differentiated between the top-cited cohorts. The respective t-statistics are plotted in Figure 5. Our results are in line with Moosa (2016) as far as the insignificance for the t-statistic for low top-cited cohorts is concerned. Moosa (2016) therefore concludes that there is no link between the mean citation

and multiple authorship. But based on our data set tests of single-authorship against three authors lead to significant positive t-statistics for all cohorts above the top 6,000, meaning that above this threshold single-authored papers obtained on average higher citation scores than papers written by three authors. Interestingly, the relationship 1 vs. 2 and 2 vs. 3 authors have positive but insignificant t-statistics, whereas the comparisons between single-, two, and three-authored papers versus papers written by four authors lead to negative t-statistics for almost all cohorts. Nevertheless, only the pairwise test of 3 vs. 4 authors is significant for a decent number of cohorts between the top 3,000 and top 8,000 papers.

These pairwise comparisons have a large disadvantage: They do not control for the factor time since publication. In the following, we account for this issue and regress the citation count on the number of citations and a time trend. We estimate the following equation

$$citations = \alpha + \beta_1 \cdot authors + \beta_2 \cdot age + \beta_3 \cdot age^2 + \epsilon \tag{1}$$

where *citations* equals the number of citations the paper received, *authors* equals the number of authors of the paper and *age* denotes the article age. In Table 2 we report the regression coefficients for the number of authors and the corresponding t-statistic. In the basic model they are negative and insignificant. The coefficients get significant only after considering more than 6,400 top-cited articles. On the contrary, taking the full sample leads to a (small) positive and significant effect of the number of authors on the number of citations of a paper. For the model with time trends, the results differ: Here, the coefficients above the top 2,200 cohort are positive, significant, and larger in size. In conclusion, there seems to be a positive effect of the number of authors on the average citation score when time effects and larger sets of top papers are taken into account. It might be that in this range the simple advantage of manpower, i.e four authors that can go to conferences, talk about their paper, and network, leads to a higher mean number of citations. In addition, like Rath and Wohlrabe (2016a) point out, more opportunities to publish working papers and higher (self-)citation numbers lead to a higher visibility of co-authored papers. Card and DellaVigna (2013) argue that researchers form groups in order to face the tougher competition and the decreasing

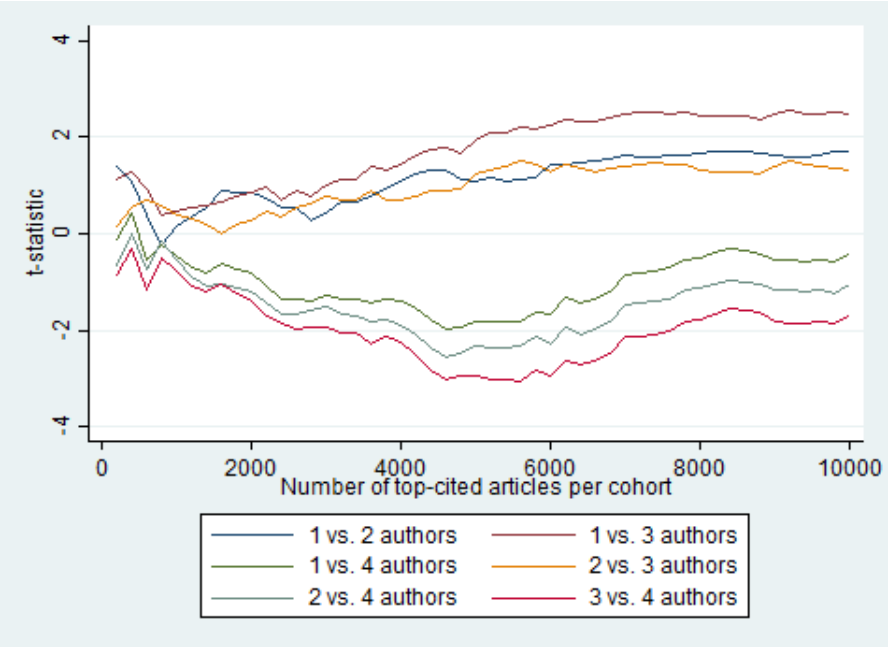
acceptance rates for top journals. Moreover, Nowell and Grijalva (2011) state that the rise in co-authorship reflects the wish for high quality, the institutional structure, and the complexity of the discipline. Hence, multi-authored papers might offer important advantages and lead to higher citation counts.

Figure 4: Citations vs. authorship

Table 2: Regression results for citations vs. authorship

	Basic Model		With time trends	
Cohort	Coefficient	t-statistic	Coefficient	t-statistic
200	-67.348	-1.051	-71.860	-0.987
400	-50.913	-1.311	-21.390	-0.500
600	-15.873	-0.549	24.534	0.795
800	-3.079	-0.135	27.965	1.192
1000	-2.876	-0.147	27.273	1.357
1200	-3.040	-0.176	26.670	1.519
1400	-2.908	-0.188	23.297	1.480
1600	-7.409	-0.567	17.241	1.266
1800	-6.705	-0.559	18.889	1.523
2000	-7.738	-0.714	17.655	1.591
2200	-6.074	-0.597	18.350	1.771
2400	-2.362	-0.246	21.966	2.239
2600	-2.967	-0.329	20.527	2.239
2800	-1.224	-0.145	21.026	2.443
3000	-4.293	-0.555	17.587	2.230
Full Sample	0.361	17.664	0.709	31.077

Figure 5: Pairwise tests of citations between various degrees of authorship: t-statistics



5 Conclusion

In this article we analyzed the links between the quality of a paper and the quality of a journal as well as the possible impact of multiple authorship on the number of citations a paper receives. We find a positive correlation between the quality of a paper, measured by the number of citations, and the quality of the journal, measured by the impact factor and the h-index. In contrast to Moosa (2016), the correlation becomes almost perfect when a larger data set is considered and when a non-linear model is used. Moreover, the relationship between the number of authors and the number of citations, which is questioned in Moosa (2016), seems to be existent in our data set. Controlling for possible time effects and including more cohorts lead to a positive influence of co-authorship on citations.

References

Arellano, M., and S. Bond (1991): "Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations," *The Review of Economic Studies*, 58(2), 277–297.

Card, D., and S. DellaVigna (2013): "Nine Facts about Top Journals in Economics," *Journal of Economic Literature*, 51(1), 144–61.

Moosa, I. A. (2016): "Citations, journal ranking and multiple authorships: evidence based on the top 300 papers in economics," *Applied Economics Letters*, p. in press.

Nowell, C., and T. Grijalva (2011): "Trends in co-authorship in economics since 1985," *Applied Economics*, 43(28), 4369–4375.

Oswald, A. J. (2007): "An Examination of the Reliability of Prestigious Scholarly Journals: Evidence and Implications for Decision-Makers," *Economica*, 74(293), 21–31.

Rath, K., and K. Wohlrabe (2016a): "Recent trends in co-authorship in economics: evidence from RePEc," *Applied Economics Letters*, 23(12), 897–902.

——— (2016b): "Trends in economics publications represented by JEL categories between 2007 and 2013," *Applied Economics Letters*, 23(9), 660–663.

Seiler, C., and K. Wohlrabe (2012): "Ranking economists on the basis of many indicators: An alternative approach using RePEc data," *Journal of Informetrics*, 6(2), 389–402.

——— (2014): "How robust are journal rankings based on the impact factor? Evidence from the economic sciences," *Journal of Informetrics*, 8(4), 904–911.

Sutter, M., and M. Kocher (2004): "Patterns of co-authorship among economics departments in the USA," *Applied Economics*, 36(4), 327–333.

Zimmermann, C. (2013): "Academic rankings with RePEc," *Econometrics*, 1(3), 249–280.