NBER WORKING PAPER SERIES

EXPLORATIONS OF CUMULATIVE ADVANTAGE
USING DATA ON FRENCH PHYSICISTS

Bronwyn H. Hall
Jacques Mairesse

Working Paper 32285
http://www.nber.org/papers/w32285

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
March 2024

This research received some financial support from the Institute of Competition and Innovation, MPI-Munich. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

NBER working papers are circulated for discussion and comment purposes. They have not been peer-reviewed or been subject to the review by the NBER Board of Directors that accompanies official NBER publications.

Explorations of Cumulative Advantage Using Data on French Physicists
Bronwyn H. Hall and Jacques Mairesse
NBER Working Paper No. 32285
March 2024
JEL No. J24,M5,O31,O38

ABSTRACT

We use a large dataset of approximately 1500 physicists employed by the Centre National de la Recherche Scientifique (CNRS) in France to investigate the role of cumulative advantage in their publication career. Measuring output by time series of the number of publications and the number of cite-weighted publications for each physicist, we find that the simple stylized facts predicted by cumulative advantage hold only weakly for these physicists. However, regressions with fixed effects, life cycle effects, and past productivity provide strong evidence that cumulative advantage matters for future productivity. In future work, we plan to compare this sample with those from other countries that have different systems for the conduct of scientific research.

Bronwyn H. Hall
University of California, Berkeley
Economics Department
123 Tamalpais Road
Berkeley, CA 94708
and NBER
bhhall@berkeley.edu

Jacques Mairesse
CREST (ParisTech-ENSAE)
5, Avenue Henri Le Chatelier
91120 PALAISEAU
FRANCE
and UNU-MERIT (Maastricht University)
and also NBER
mairesse@ensae.fr

Explorations of cumulative advantage using data on French physicists

Bronwyn H. Hall and Jacques Mairesse

I. Introduction

The humorous quotation from David Lodge's novel "*Small World*", in the footnote below,[4] describes the phenomenon in scientific research that was termed the "Matthew Effect" by Merton (1968). It provides an explanation for the empirical observations by Lotka (1926) and later researchers that the majority of papers in a discipline are produced by a small minority of scientists in that discipline. But there is another possible explanation of extreme skewness in scientific production: it may simply reflect skewness in the intrinsic productivity of individual scientists, that is, skewness in "fixed effects". For example, Huber (1998) found the fixed effect model to be a sufficient explanation for the distribution of inventor productivity in the form of patents.

There is a long history of econometric and statistical research that attempts to distinguish these two interpretations of the evolution of individual outcomes over time, often assuming discrete data, most commonly of the binary (0-1) form. See Bates and Neyman (1951) for application to accident data, Heckman (1978, 1981a, 1981b) for labor force participation, and Halliday (2007) for health status data. In these applications, the question is whether the outcomes can be explained by intrinsic individual heterogeneity or whether there is also state dependence in the sense that future outcomes depend on earlier outcomes in ways that cannot be explained solely by heterogeneity. In our application, cumulative advantage is equivalent to observing state dependence in the data.

In this paper we use a new dataset on publications by physicists from Institute of Physics in the French CNRS (Centre National de la Recherche Scientifique) to explore the extent to which the phenomenon of cumulative advantage is needed to explain the evolution of publications, both counts and citation-weighted counts. Do outcomes during the career of

[4] Morris Zapp in David Lodge: *Small World* (Penguin Edition, page 151), contemplates his impending visit to the Rockefeller Foundation at the Villa Serbelloni in Bellagio: "The beauty of academic life is that to them that had had, more would be given…… All you needed to do to get started was to write one really damned good book, which admittedly wasn't easy when you were a young college teacher just beginning your career …… But on the strength of that one damned good book you could get a grant to write a second book in more favorable circumstances; with two books you got promotion, a lighter teaching load, and courses of your own devising; you could then use your teaching as a way of doing research for your next book, which you were thus able to produce all the more quickly. This productivity made you eligible for tenure, further promotion, more generous and prestigious research grants, more relief from routine teaching and administration. In theory, it was possible to wind up being full professor while doing nothing except to be permanently absent on some kind of sabbatical grant or fellowship."

these physicists influence their subsequent success, or is the observed heterogeneity simply explainable by variations in intrinsic productivity. Intuitively, this will amount to asking whether publications are random across time (controlling for research age) or whether the order of the time series for each individual matters in predicting their outcomes.

The stylized empirical facts about scientific productivity can be summarized as follows:

1. The distribution of individual scientific productivity is extremely dispersed, as reflected by the fact that most papers, as well as the most highly cited papers, are published by a small share of scientists. (Lotka 1926)
2. The resulting publication and citation hierarchy among scientists remains very stable during much of their lifetime (David 1994)
3. The corresponding concentration tends to increase over time within scientific cohorts. (Stephan 1996)

These stylized facts have been interpreted by Merton (among others) as indications of "cumulative advantage" or the "Matthew Effect" in the publication of research papers. Success in publishing early in one's career, especially publishing of highly cited papers, is presumed to lead to greater resources and therefore greater publication and citation rates than other researchers as the career moves forward. A number of empirical studies have investigated this question in the past, mostly with much smaller datasets than ours (for example, Allison et al. 1982; Price 1976).

The previous work has mostly confirmed the stylized facts above; unfortunately, confirmation of those facts is not really enough to confirm the presence of cumulative advantage, as wide variations in intrinsic productivity can also generate data that displays these characteristics. The problem is compounded by the fact that the leading parametric model used to describe the data is the negative binomial, which can be generated by either heterogeneity or state dependence.

II. The Institute of Physics at CNRS

Our sample is a large group of physicists affiliated to the Institute of Physics (INP) in France, one of the prominent Institutes of CNRS. INP is part of one large interdisciplinary public research organization under the responsibility of the French Ministry of Education and Research, the National Centre for Science Research (CNRS). CNRS is one of the most prestigious French research institutes, created by the state in the 1940s with the mission of 'advancing knowledge for the benefit of society,' while 'respecting ethical rules and showing commitment to professional equality.'[5]

[5] http://www.cnrs.fr/fr/le-cnrs.

In this section we briefly describe the features of this institute and the research careers that it enables. France has a reputation for producing cutting-edge research in the field. Looking back in history, important discoveries have been attributed to French scientists. For instance, the international system of measurement units of electric current, i.e. the ampere, was introduced by André-Marie Ampère (1775-1836) who was one of the founders of the science of electromagnetism. Several discoveries have been awarded the highest worldwide reward, the Nobel Prize. For example, in 1903 Marie Curie (1867-1934) was awarded the prize for her pioneering research on radioactivity; in 2012 Serge Haroche was awarded the prize for implementing innovative experimental methods of measuring and manipulating individual quantum systems; and in 2018, Gérard Albert Mourou was awarded the prize for his discoveries on very high-intensity laser pulses

CNRS researchers in general and INP scientists in particular are selected through an extremely competitive process, one that is much more competitive than that for French universities, even the most selective ones. Once selected, they can generally not be fired and have a job for life. Researchers affiliated to INP are French civil servants and follow a well-defined career progression with three career levels. Specifically, a researcher enters INP as *Chargé de Recherche 1* (CR1), and later usually upgraded to *Chargé de Recherche 2* (CR2); and, if as CR he is recognized for outstanding academic achievements, he can be promoted to *Directeur de Recherche* (DR). In fact, a large number of CNRS researchers remain at the (CR)s and are never promoted to (DR)s. Researchers are appointed according to their expertise, and, in each career level, they have different responsibilities. (CR)s are generally part of a lab directed by a (DR), where they work under his supervision, usually on the same research projects, sometimes on personal independent research project.

INP scientists have a high academic profile and productivity. Publication data retrieved from the Web of Science (Clarivate Analytics) shows that during the period 2001–2016, they produced approximately 52,000 publications in physics.[6]

III. Data and selectivity

The source of our data are two panel datasets supplied by the CNRS at two different points in time, 2005 and 2015, containing a total of 1,490 physicists. Each contained the full publishing career to date of the physicists present in that year, as well as their gender, birth year, and rank (*Charge de Recherche* or *Directeur de Recherche*, roughly corresponding to assistant and full professor ranks at the university). Earlier versions of these data have already been analyzed in other ways by Mairesse and Pezzoni (2015) and Mairesse et al. (2019). We have updated them by extracting the yearly paper counts and their citations from

[6] Average salaries of (CR)s and (DR)s currently vary from 2,000 to 6000 euros per month.

Open Alex as described in the appendix. Thus the sampling frame is simply a dataset containing the population of CNRS physicists with their names, birth years, and genders, which we take to Open Alex to obtain publication data (Priem et al. 2022).

After cleaning, the resulting panel dataset contains data on 1,439 physicists born between 1938 and 1986, with up to 64 years of publication information for each, although the average number of years of data for each physicist is 38 years. A probit regression for the probability of being in the sample reveals that the missing 51 physicists are randomly distributed across birth year and gender. We also created a balanced panel that contains 888 physicists born between 1938 and 1968, each with 35 years of data (ages 21 to 55). Further information on the data is supplied in Appendix A.

We first examine the characteristics of the physicists that exit our sample before the cutoff date of 2022. Note that exit is defined by the date after which the physicist no longer publishes. There are 494 such physicists; the remaining 945 are censored at 2022. We expect that those that exit earlier will have sparser publication histories than those which remain. To explore this question, we estimated a hazard rate model for exit has a function of the physicist's birth year, their gender, and the number of papers or cites accumulated prior to our cutoff date to the papers published before exit. The results are shown in Table 1, for a simple Cox proportional hazard model, a Weibull model, and a Weibull model that allows for a normally distributed random effect in the hazard rate. The Weibull model shape parameter estimate (ln_p) clearly indicates increasing hazard over time and the estimated probability curve shows sharply increasing exit probabilities between ages 56 and 67.

All the columns of the table show the same result. Exit rises with later birth years, and is substantially higher for female physicists. As expected, the most important predictor of exit is the past publication history, whether measured as publication counts or as cite-weighted publication counts. The odds ratios for either are well below unity and highly significant, indicating that exit is much less likely for those physicists that have a greater number of prior publications. Interestingly, the number of publications is a much stronger predictor than the cite-weighted number. The results support the idea that one effect of cumulative advantage in research is that it encourages a successful researcher to remain active.

Table 1

Hazard rate estimation of exit probability

Method	*Cox*	*Weibull*	*Weibull RE*	*Weibull RE*	*Weibull RE*
Birth year	1.158***	1.146***	1.146***	1.168***	1.168***
	(0.012)	(0.013)	(0.013)	(0.018)	(0.018)
Gender	1.454***	1.429***	1.429***	1.427***	1.429***
	(0.170)	(0.163)	(0.163)	(0.177)	(0.177)
Log (cum. cites)	0.651***	0.647***	0.647***		1.022
lagged	(0.023)	(0.026)	(0.026)		(0.080)
Log (cum. pubs)				0.441***	0.430***
lagged				(0.050)	(0.063)
ln_p (shape)		9.224***	9.224***	10.378***	10.386***
		(0.524)	(0.524)	(0.813)	(0.815)
Chisquared	304.0	226.8	226.8	101.1	101.4
Degrees of freedom	3	3	3	3	4
Log likelihood	-2,731.2	-175.6	-2,002.4	-1,968.5	-1,968.5

53,670 observations on 1,439 physicists. 494 failures and 945 censored.

Standard errors are robust and clustered on physicists.

Hazard ratios and their standard errors are shown.

IV. Do the data exhibit the expected behavior?

To explore the extent to which these data exhibit the behavior associated with cumulative advantage rather than being simply a consequence of variations in individual productivity that does not evolve over time, we conducted an exploration in the spirit of Huber (1998), who found that the individual distributions of inventor patenting over time were adequately described by a Poisson model without overdispersion. Our exploration is in Appendix B, where we conclude that Poisson for each physicist's paper production along with a gamma distribution for their average productivity differences is a pretty good description of our publication data. However, especially for those with larger numbers of papers and for citation-weighted paper counts, the Poisson distribution with a fixed productivity for each physicists is rejected in favor of the negative binomial or other model with higher variance than the Poisson, which means that we cannot rule out some role for cumulative advantage (Allison 1980). This result is confirmed by simple parametric fits to the data in Appendix C, which show that although the negative binomial model comes the closest, the publication data are more dispersed across physicists than suggested by that model.

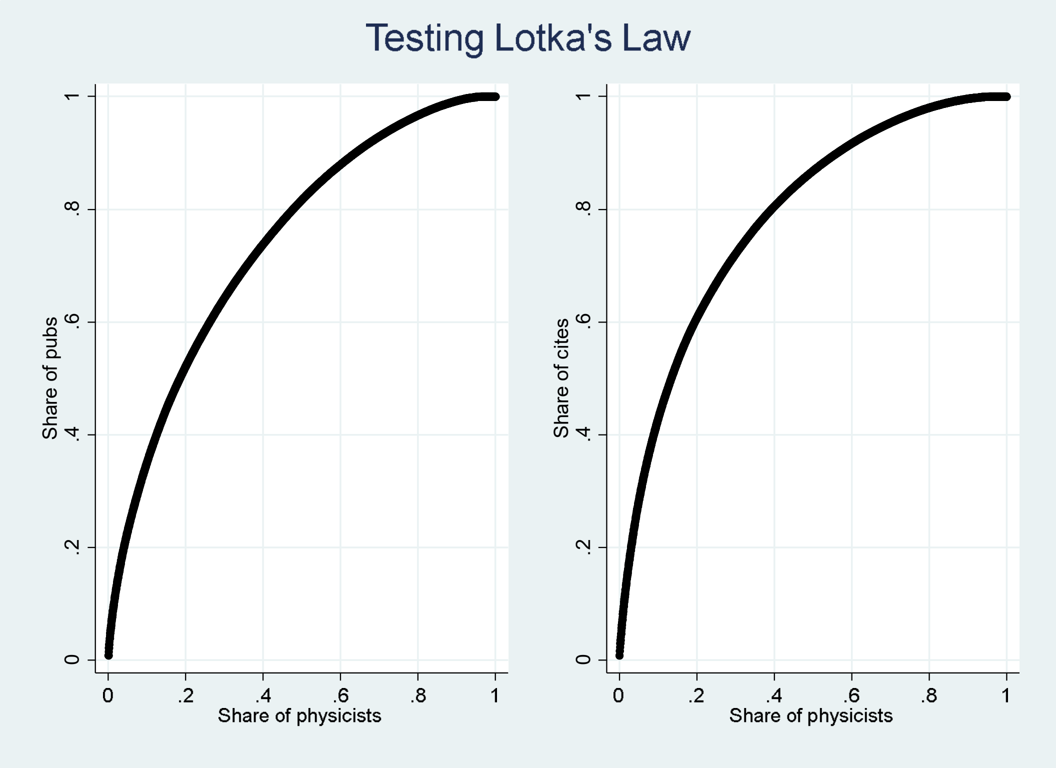
In what follows, we first explore the ways in which the data exhibited the behavior described in the three points in the introduction: dispersion across scientists, stable hierarchies, and increasing concentration. Then we use some "runs" tests to test for cumulative advantage.

1. Dispersion

Proposition 1 in the introduction implies both a quantity and quality dimension to the dispersion of physicist publications. Loosely speaking, Lotka's Law states that 20 per cent of the physicists should account for 80 per cent of the output. For quantity we use the total number of publications for each physicist and for quality we use the sum of the citation counts for these publications.

Figure 1 displays the graphs of the shares of publications and citations ranked from high to low versus the share of physicists that account for them. Clearly they do not satisfy the informal version of Lotka's Law: roughly 54 per cent of publications and 61 per cent of citations are contributed by 20 per cent of the physicists. So there is some dispersion in output, but not as extreme as suggested by Lotka. This is perhaps to be expected, given the selectivity of CNRS physicists in the first place.

Figure 1

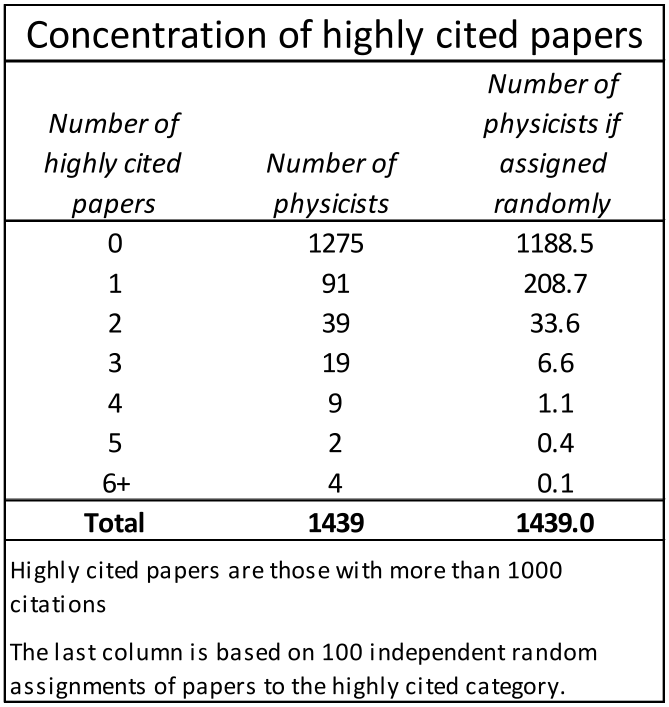


We can also ask whether the most important papers are concentrated among a few physicists. To do this, we compute the number of physicists that have produced papers cited more than 1000 times. There are 303 such papers, or 0.11%, produced by 164 out of 1,439 physicists (11.5%). To compare the distribution to randomly occurring highly cited papers, we assign each of the approximately 265,000 papers randomly to the highly cited status and compute the number of physicists with such papers, repeating this 100 times and averaging.

This allows us to compute the expected distribution of highly cited papers under the hypothesis of equal probability across papers. We can then perform a chi-squared test for the hypothesis that highly cited papers are distributed randomly across physicists.

The results are shown in Table 2, where the second column is the observed distribution and the third column is the predicted number if papers are randomly highly cited. The value of the chi-squared test for equality of the columns is 267.4 with 6 degrees of freedom, which clearly rejects randomness of assignment to a physicist. If we define highly cited as more than 500 citations instead, there are 1,219 (0.46%) highly cited papers and the chi-squared is 333.5 with 15 degrees of freedom (not shown). So the hypothesis that highly cited papers are random across physicists is clearly rejected, implying some concentration of highly cited papers on a few physicists. Interestingly, the discrepancy is due to two things: more physicists have no highly cited papers, fewer have one highly cited paper, but at the same time, there are more physicists with 3 or more highly cited papers (34 instead of 8.2) than predicted.

Table 2



Concentration of highly cited papers

Number of highly cited papers	*Number of physicists*	*Number of physicists if assigned randomly*
0	1275	1188.5
1	91	208.7
2	39	33.6
3	19	6.6
4	9	1.1
5	2	0.4
6+	4	0.1
Total	**1439**	**1439.0**

Highly cited papers are those with more than 1000 citations

The last column is based on 100 independent random assignments of papers to the highly cited category.

In Appendix C, we explore the functional form of the publication and citation distributions, using the Pareto and log normal distributions that are commonly used to model skew distributions in the innovation literature (e.g., Scherer, 1998) as well as the Poisson and negative binomial models commonly used for discrete data (Hausman et al, 1984). For both publications and citation-weighted publications, the estimated Pareto parameter is substantially less than one, which suggests that neither distribution has a mean or a variance.

However, the Pareto shape was a poor fit to the data, because it predicts far more concentration than actually exists in the data.

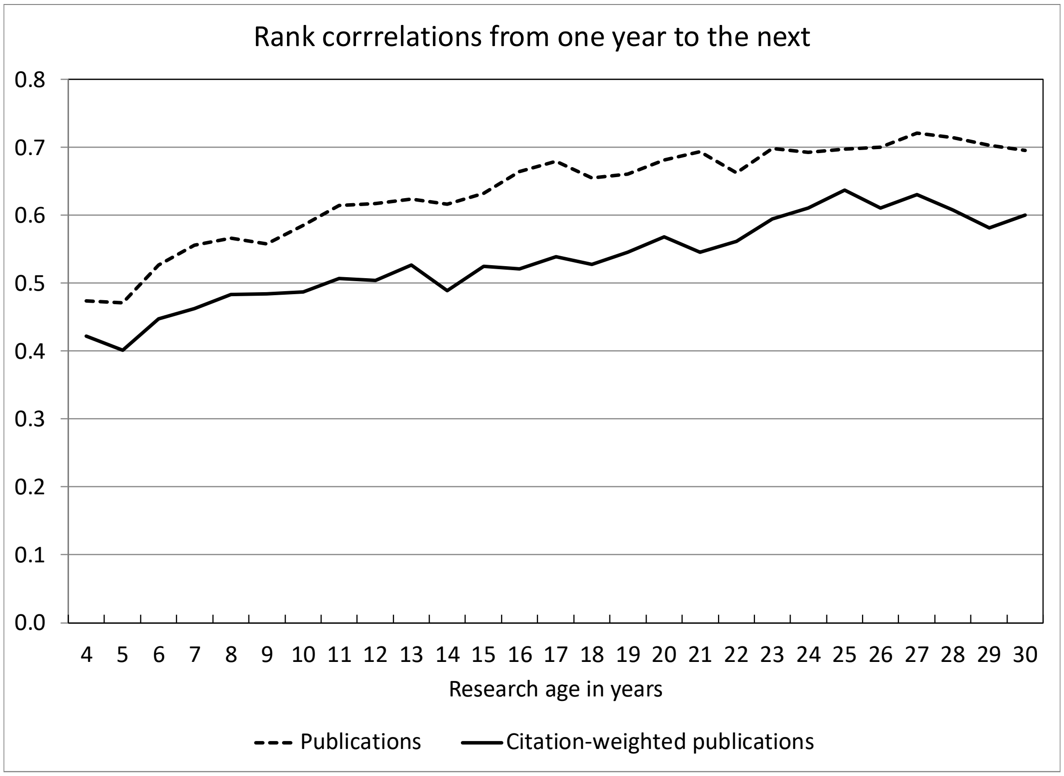
By comparing the cumulative distribution predicted by these models to the actual, we found that the negative binomial model clearly provided the best fit for these data, somewhat better for the publication counts than for the citation weighted publication counts. In Appendix B we provide results for estimation of the separate ingredients of the panel negative binomial: Poisson for the individual physicist and gamma for the average productivities across physicists. We found that although the Poisson was clearly rejected for most of the individual physicists in favor of a negative binomial, the negative binomial provided a reasonably good fit to their average productivity distribution, albeit one that had a lower variance than the data.

Our conclusion is that there is support for dispersion in publication across these physicists, but that it is not as extreme as implied by Lotka's Law, and that the negative binomial is a pretty good description of the publication data, but not when the publications are cite-weighted.

2. Stability

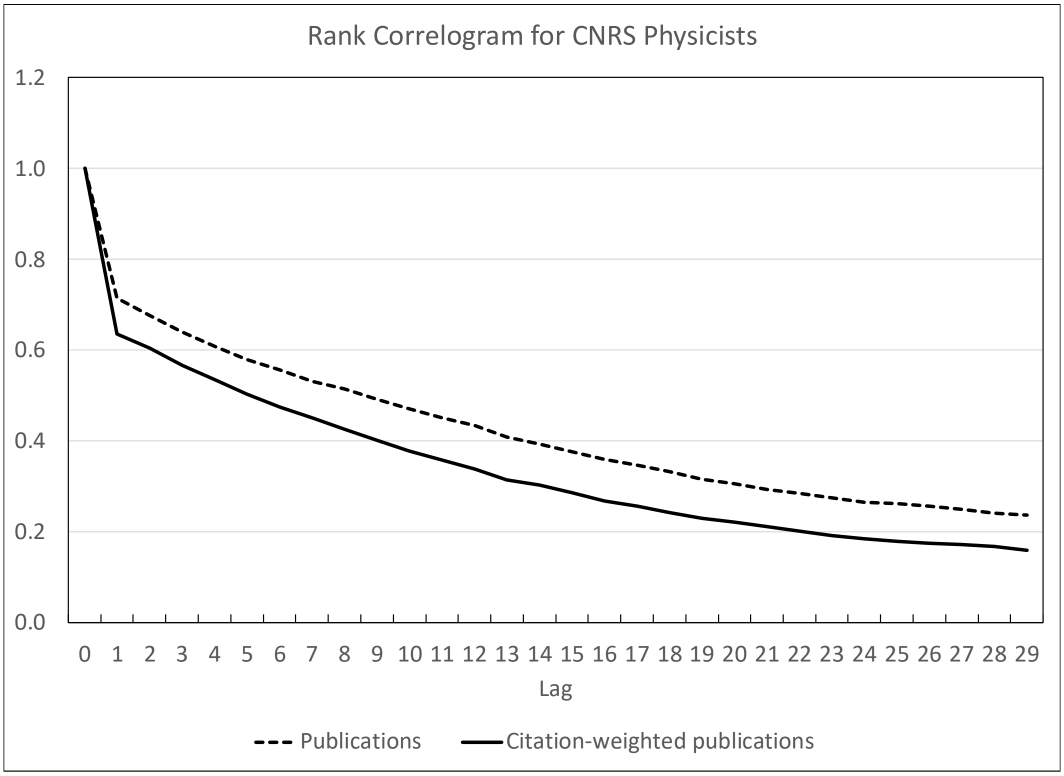
To check whether the publication hierarchy among physicists remains stable during their career, we compare the correlations of their rank among physicists with the same research age over time. The result is shown in Figure 2 for the balanced panel of 896 physicists. This figure shows that year-to-year, the hierarchy of physicists is relatively stable, and that stability flattens out but persists about 20 years into their careers.

Figure 2



However, when we look at the full correlogram for the ranks, we find in Figure 3 that the hierarchy is not stable over the entire career. The correlation of ranks falls as the lag between the observations increases. By lag 11, the correlation across years of the ranks of both publications and citation-weighted publications has fallen below 0.4. So we cannot really say that the hierarchy of physicists as measured by publications is stable throughout their career, although it is moderately persistent from year to year. In addition, the citation-weighted publications are somewhat less stable than the publication counts.

Figure 3



3. Increasing concentration

Stylized fact 3 in the introduction predicts that the concentration of output for a given group of scientists tends to increase over time. We explore whether this holds in our data using a conventional measure of inequality, the Gini coefficient and Lorenz plots. We show the evolution over time of the Gini coefficients for publications and impact factor-weighted publications in Figure 4. Research age is defined as the time since the physicist published their first paper.[7] For publications, concentration measured by the Gini is slightly noisy in the first couple of years and is then flat. The Gini for cite-weighted publications declines until about research age 18 and then is mostly flat throughout the research career. Both show a very slight increasing tendency after research age 20.

[7] The initial research age varies from age 21 to age 45, which means that the Gini in the later years is based on fewer observations than that for research ages of 20 or less. However, using the more limited balanced panel and setting the starting age to 21 for all physicists produces roughly the same conclusion, a flat Gini for publications and cite-weighted publications.

Figure 4

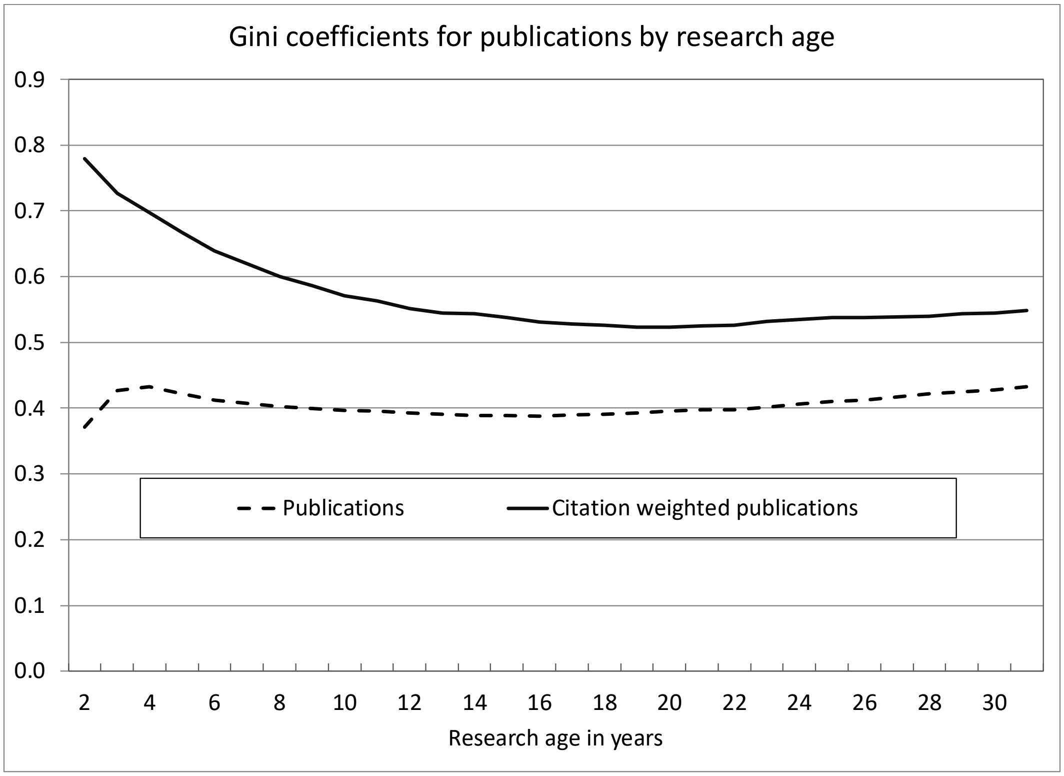
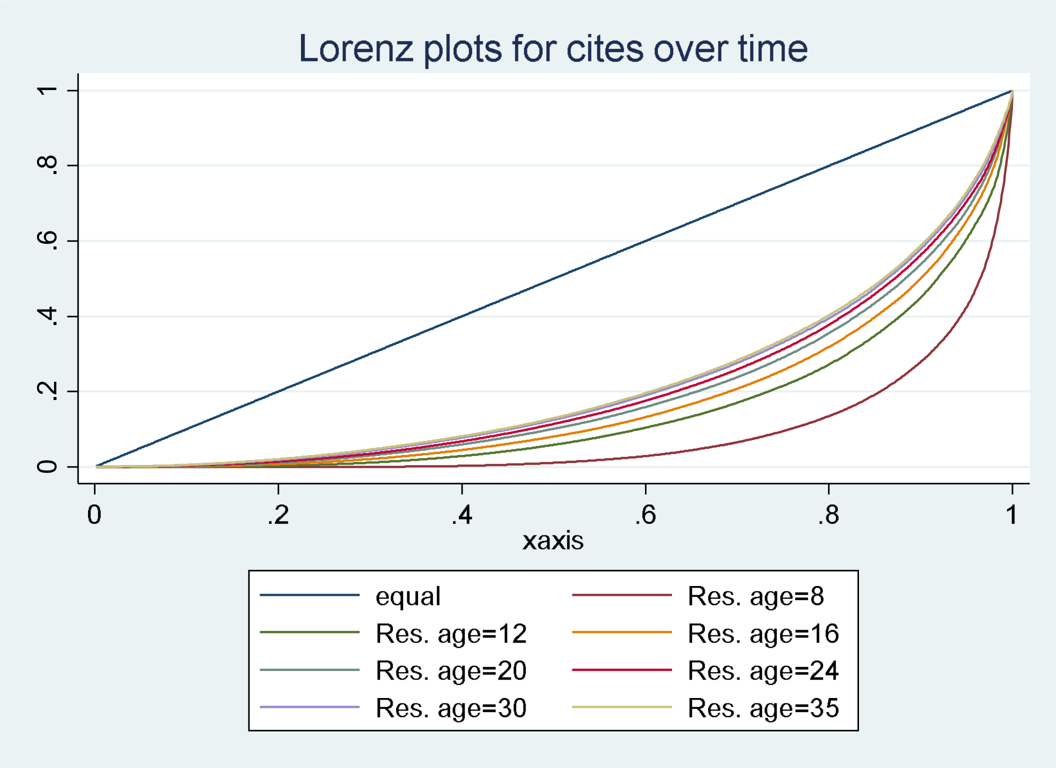


Figure 5 shows Lorenz curves for the balanced sample at different ages for citation-weighted publications (unweighted publications look much the same). As the research age increases, the Lorenz curve moves closer to equality and looks like it is converging, with ages 24, 30, and age 35 almost coinciding. So the distribution of physicist productivity is unequal but does not become greater over time, contradicting stylized fact 3 (increasing concentration).

Figure 5

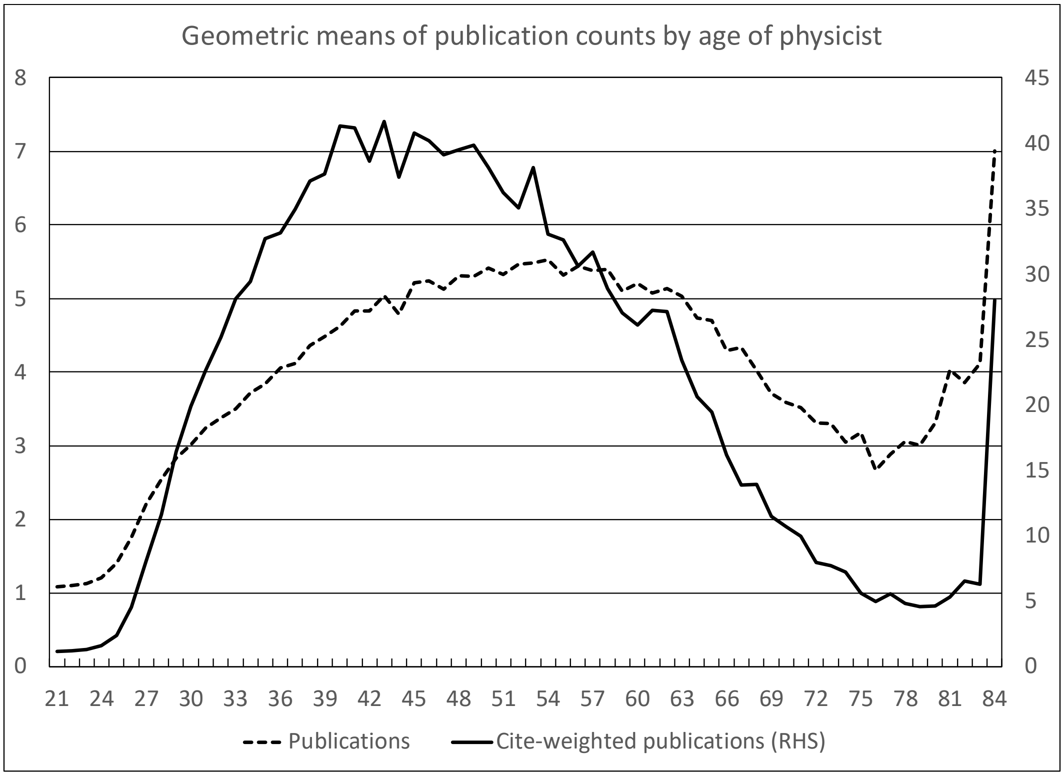


V. Regression estimates

To look more closely for evidence of cumulative advantage in future productivity in the presence of heterogeneity, we estimated a series of simple panel regression models for publications and cite-weighted publications that included both fixed effects and lagged cumulative productivity. The models were estimated using the entire sample of physicists and years, for ages 21 to the last age of publication for each physicist.

There is a life cycle effect in the publication series for most researchers, where publication rates increase at an increasingly slower rate with age, eventually peaking sometime during the 40s or 50s (e.g., Hall et al. 2007). In Figure 6 we show the geometric means of the two publication counts across all physicists in the sample as a function of their age. Both show this pattern with publication counts apparently peaking in the early 50s, while cite-weighted counts peak earlier, due to the truncation of cites for later papers. Note that the numbers are based on fewer physicists at the later ages, both due to truncation and because they age out. At age 84, there is only one (highly productive) physicist remaining in the sample, hence the spike at that age.

Figure 6



It is well-known that linear models including dummy variable models that control for age, cohort, and time period are not identified because in most settings, age is a simple linear function of cohort and the current period. Hall et al. (2007) investigated the interpretation of regressions based on scientist publication data that attempt to include full sets of dummies for age, cohort, and year. Note that when fixed individual effects are included in a regression, they will subsume the cohort effects, so full age and year effects will not be identified. Our approach to identification here is twofold: first, we use a quadratic in age with and without a full set of year dummies. Second, we estimate while including either random age and age squared coefficients across researchers or a cohort specific quadratic in age.

The models we consider are the following:[8]

1. Panel random effects with the log of the counts as the dependent variable.
2. Panel fixed effects with the log of the counts as the dependent variable.
3. A random coefficient version of OLS where each physicist has his or her own set of coefficients, and they are assumed to be drawn from a distribution with mean and variance. (Swamy 1970)
4. Poisson fixed effect model for the counts.
5. Negative binomial fixed effect model for the counts.

[8] Note that an investigation of the functional form of the relationship between current publications and past cumulative publications revealed that the relationship was approximately linear in the log, as specified here.

The results are shown in Table 3. Note that the coefficient of gender is not identified when we use fixed effects, except in the negative binomial case. For publications, gender is rather insignificant, whereas females have lightly higher cite-weighted publications, unless year is controlled for. This is presumably because the number of women in the sample increases over time and those who entered early are selected to be more productive. The Hausman test for models 1 and 2, clearly rejects the random effects model in favor of one with fixed effects. With the exception of the random coefficient model, all the models deliver the same qualitative and quantitative result: lagged cumulative productivity is strongly positively related to current productivity, above and beyond any average fixed difference across the physicists. A doubling of lagged cumulative publications leads to fifty per cent more publications in the current year, while a doubling of lagged cite-weighted publications leads to one-third more cite-weighted publications.

The fixed effect models in columns 2, 4, 5, and 6 have a low peak productivity age of about 20-25, and a substantial role for lagged productivity. Clearly lagged productivity can account for a great deal of the age-publication profile, leading to somewhat nonsensical peak age effects. Model 6 shows that adding year effects makes little difference to the estimates once we control for lagged productivity and fixed individual effects.

The exception to these estimates is the random coefficient model. For this model, each physicist is allowed to have his own age-publication profile, but the estimates are averaged in such a way that those with high variance have lower weight. So the model is somewhat more flexible at the physicist level. These estimates have peak productivity ages that are closer to those in Figure 6, and a somewhat weaker cumulative advantage effect, presumably because average productivity for each individual is imperfectly controlled for.

The final column shows the results for our balanced panel of 886 physicists and the negative binomial fixed effects sample. Although the estimates for this sample differ slightly quantitatively, they are qualitatively similar to those for the whole sample and confirm a substantial dependence of current productivity on past productivity in the presence of fixed productivity effects.

Table 3

Regression Estimates

	(1)	*(2)*	*(3)*	*(4)*	*(5)*	*(6)*	*(7)*
Method	*RE*	*FE*	*Random coeff.*	*Poisson FE*	*Neg. Bin. FE*	*Neg. Bin. FE*	*Neg. Bin. FE*
Dep. Var.	*Log count*	*Log count*	*Log count*	*Count*	*Count*	*Count*	*Count (bal.)*
				Publications			
Age	-0.0095**	0.0133***	0.2147***	0.0019	0.0276***	0.0142***	0.1115***
	(0.0039)	(0.0043)	(0.0105)	(0.0095)	(0.0032)	(0.0033)	(0.0074)
Age squared	-0.0002***	-0.0004***	-0.0025***	-0.0004***	-0.0006***	-0.0006***	-0.0015***
	0.0000	0.0000	(0.0001)	(0.0001)	0.0000	0.0000	(0.0001)
Peak age (derived)	-23.6	17.8	42.3	2.6	22.6	12.7	36.9
	(14.1)	(3.8)	(0.7)	(12.5)	(1.6)	(2.3)	(0.7)
Log (cum. pubs) lagged	0.4850***	0.4230***	0.1158***	0.6153***	0.5586***	0.5136***	0.5129***
	(0.0067)	(0.0080)	(0.0141)	(0.0194)	(0.0075)	(0.0076)	(0.0112)
Gender	-0.0585***				0.0295	-0.0246	0.0610
	(0.0174)				(0.0309)	(0.0318)	(0.0448)
Year dummies	no	no	no	no	no	yes	no
Log likelihood		-49,225.9		-130,605.1	-110,728.6	-109,856.0	-57,143.2
Chi-squared	13,614.4		2,572.5	4,844.1	20,984.8	.	13,146.4
Degrees of freedom	4	2	3	3	4	32	4
				Cite-weighted publications			
Age	0.0428***	0.0798***	0.6980***	0.0156	0.0357***	0.0290***	0.1231***
	(0.0085)	(0.0093)	(0.0313)	(0.0173)	(0.0034)	(0.0034)	(0.0085)
Age squared	-0.0010***	-0.0013***	-0.0091***	-0.0006***	-0.0007***	-0.0006***	-0.0017***
	(0.0001)	(0.0001)	(0.0005)	(0.0002)	0.0000	0.0000	(0.0001)
Peak age (derived)	20.6	29.6	38.2	12.5	24.9	22.7	35.7
	(2.4)	(1.5)	(0.6)	(10.4)	(1.2)	(1.5)	(0.5)
Log (cum. cites) lagged	0.5333***	0.4806***	0.1681***	0.3599***	0.3536***	0.3344***	0.3438***
	(0.0074)	(0.0089)	(0.0174)	(0.0205)	(0.0037)	(0.0039)	(0.0052)
Gender	-0.1542***				0.0424**	0.0186	0.0952***
	(0.0442)				(0.0177)	(0.0179)	(0.0247)
Year dummies	no	no	no	no	no	yes	no
Log likelihood		-102,289.2		-4,774,667.9	-233,153.5	-232,204.2	-124,106.2
Chi-squared	13,149.0		2,054.1	757.2	17,159.7	.	10,169.0
Degrees of freedom	4	2	3	3	4	15	4
Number of observations	53,555	53,555	53,555	53,555	53,555	53,555	30,158
Number of physicists	1,435	1,435	1,435	1,435	1,435	1,435	887

Standard errors are robust and clustered on physicists.

The last column is for balanced data ages 21-55 only.

The random coefficient results in Table 3 both suggest that the age-productivity relationship is different for different physicists. In particular, we know that the citation counts will be lower for physicists with more recent birth dates, and that they will peak sooner, due to the truncation at 2023. This prediction was confirmed by plotting the age-productivity distribution by birth year (not shown). To accommodate this variation, we allowed the age and age squared coefficients to vary by birth year (cohort) in the balanced sample regressions shown in Table 4.[9]

Table 4

Regression Estimates with separate age coefficients by birth year

	(1)	*(2)*	*(3)*	*(4)*	*(5)*	*(6)*	*(7)*
Method	*RE*	*FE*	*Poisson RE*	*Poisson FE*	*Neg. Bin. RE*	*Neg. Bin. FE*	*Neg. Bin. FE*
Dep. Var.	*Log count*	*Log count*	*Count*	*Count*	*Count*	*Count*	*Count*
				Publications			
Log (cum. pubs) lagged	0.4733***	0.4133***	0.6082***	0.5435***	0.6425***	0.5059***	0.4835***
	(0.0085)	(0.0099)	(0.1106)	(0.0232)	(0.0110)	(0.0117)	(0.0118)
Gender	-0.0487**		-0.0874*		-0.0141	0.0524	0.0241
	(0.0246)		(0.0501)		(0.0301)	(0.0464)	(0.0466)
Year dummies	no	no	no	no	no	no	yes
Hausman test FE vs RE (d.f.)	NA		38.2 (30)		NA		
Log likelihood		-26,772.7	-70,260.5	-65,396.2	-61,736.8	-56,971.3	-56,716.9
Chi-squared	10,010.2		14,080.2	5,831.7	14,531.6	13,431.5	.
Degrees of freedom	60	58	60	59	60	60	79
				Cite-weighted publications			
Log (cum. cites) lagged	0.4740***	0.3908***	0.2595***	0.2594***	0.3703***	0.3312***	0.3188***
	(0.0106)	(0.0129)	(0.0334)	(0.0334)	(0.0054)	(0.0055)	(0.0056)
Gender	-0.2026***		-0.0688		0.0314	0.0661***	0.0555**
	(0.0683)		(0.5711)		(0.0234)	(0.0254)	(0.0254)
Year dummies	no	no	no	no	no	no	yes
Hausman test FE vs RE (d.f.)	193.6 (57)***		0.6 (33)		2007.3 (60)***		
Log likelihood		-57,099.0	-2,621,178.9	-2,612,211.2	-131,798.3	-123,828.6	-123,690.8
Chi-squared	8,606.7		4,037.9	1,666.9	11,507.0	10,622.0	.
Degrees of freedom	60	58	60	59	60	60	112

Standard errors are robust and clustered on physicists.
Each set of estimates includes age and age-squared interacted with a complete set of dummies for the birth year of the physicist.
30,124 observations on 886 CNRS physicists aged 21-55.

Table 4 shows linear regression, Poisson, and negative binomial estimates for the model with lagged cumulative counts of publications or citations, gender when identified, and birth year-specific age and age squared, both with random physicist effects and fixed effects. We conducted Hausman tests for random versus fixed effects, although not all of them produced valid estimates in our sample. The main conclusion from these tests was that random effects were valid for publication counts, but rejected soundly for cite-weighted publications.

Allowing for birth-year specific age profiles makes little difference to the estimates of cumulative advantage. The fixed effects negative binomial estimates for publications show an elasticity of 0.51 with respect to lagged publications, while the fixed effects negative

[9] We combined birth years 1938-1940 because they were sparse. Therefore there were 29 birth years (1940-1968) with 2 coefficients each, for an additional 58 coefficients.

binomial estimates for cite-weighted publications have an elasticity of 0.33 with respect to lagged citations. Thus we can conclude from these basic regressions that there is evidence that past publication success contributes to future success, controlling for overall differences in researcher productivity and birth-year specific age-productivity profiles.

VI. Conclusions

In this paper we investigated several of the predictions of cumulative advantage in scientific research productivity using a larger set of data than most prior investigations in this area. We first looked at the evidence for three stylized characterizations of scientist productivity: Lotka's Law, the stability of ranks over time, and the increased concentration of output over the careers of researchers. We found only limited support for these predictions among our CNRS physicists. Typically only 50-60 per cent of output was contributed by 20 per cent of physicists rather than the 80 per cent predicted by Lotka's Law.

Although the correlation of physicist rankings from year to year was fairly high, reaching 0.6-0.7 at the end of their careers, it was by no means constant over the whole career, yielding mixed results for stability of the rankings. Finally, the concentration of output across physicists did not increase as predicted, but was roughly constant, with a Gini of 0.5 for publications and 0.68 for cite-weighted publications.

Although the evidence for cumulative advantage was fairly weak when we looked at the stylized facts often predicted, regressions that included lagged productivity (publications or cite-weighted publications) produced strong evidence that cumulative advantage was present even when we included fixed effects that allowed for each physicist to have his or her own average productivity. In fact, when we allowed a separate age quadratic for each birth year cohort, a Hausman test accepted random effects in average productivity, which implies that average productivity differences are random and unrelated to past productivity. In other words, cumulative advantage is more important in generating the observed differences in publication productivity over the career.

However, we caution that CNRS physicists are a selected group, with continuous research support that does not depend on the repeated acquisition of grant funds over the career. Therefore our findings may not generalize to other settings, such as university researchers in the United States. In future work, we hope to compare our sample here with physicists in Germany and the US, in an effort to learn something about the productivity of different systems of supporting basic research and science.

References

Allison, P.D., Long, J.S., and Krauze, T.K. (1982), Cumulative Advantage and Inequality in Science, *American Sociological Review* 47(5): 615–625.

Bates, G., and Neyman, J. (1951). Contributions to the Theory of Accident Proneness II: True or False Contagion. University of California Publications in Statistics 1 (1951): 215-53.

Blundell, R., R. Griffith and F. Windmeijer (2002). Individual effects and dynamics in count data models. *Journal of Econometrics* 108(1): 113-131.

Blundell R. and Preston, I. (1998), Consumption Inequality and Income Uncertainty, *Quarterly Journal of Economics* 113(2): 603-640.

Cole, J.R., and Zuckerman, H. (1984), The productivity puzzle: Persistence and Change in Patterns of Publication of Men and Women Scientists. In P. Maehr & M. W. Steinkamp (Eds.), *Advances in Motivation and Achievement*, 217-256. Greenwich, CT: JAI Press.

David, P.A. (1994), Positive Feedbacks and Research Productivity in Science: Reopening Another Black Box. In O. Grandstrand (Ed.), *Economics and Technology*, 65–89.

Eggenburger, F., and G. Polya (1923). Über die Statistik verketteter Vorgänge. *Zeitschrift für Angewandte Mathematik und Mechanik*, Volume 3, 279-289.

Ejrnæs, M. and M. Browning (2014), The Persistent–Transitory Representation For Earnings Processes. *Quantitative Economics* 5(3): 555-581.

Feller, W. (1950), *An Introduction to Probability Theory and its Applications*, Volume I. New York, NY: John Wiley and Sons (3rd edition).

Hall, B. H. (2008), Discussion of "Citation Hierarchies among Scientific Journals," EPFL conference on science and scientists (September 2008).

Hall, B. H. (1987), The Relationship between Firm Size and Firm Growth in the Us Manufacturing Sector. *Journal of Industrial Economics* 35: 583-606.

Hall, B. H. and C. Cummins (2017). *TSP 5.1 Econometric software*. Available at https://sites.google.com/view/clintcummins/tsp-5-1?pli=1

Hall, B. H., J. Mairesse, and L. Turner (2007), Identifying Age, Cohort and Period Effects in Scientific Research Productivity: Discussion and Illustration Using Simulated and Actual Data on French Physicists, *Economics of Innovation and New Technology* 16(2): 159–177.

Hall, R. E. (1978), Stochastic implications of the life-cycle permanent income hypothesis: theory and evidence, *Journal of Political Economy* 78(6): 971-987.

Hall, R. E. and F. S. Mishkin (1982), The Sensitivity of Consumption to Transitory Income: Estimates from Panel Data on Households. *Econometrica* 50(2): 461-481.

Halliday, T. J. (2007).. Testing for State Dependence with Time-Variant Transition Probabilities. *Econometric Reviews* 26(6), 685-703.

Hammermesh, D. S. (2018), Citations in Economics: Measurement, Uses and Impacts, *Journal of Economic Literature* 56 (1): 115-156.

Hausman, J. A., B. H. Hall, and Z. Griliches (1984). Econometric Models for Count Data with an Application to the Patents-R&D Relationship. *Econometrica 52*: 909-37.

Heckman, J. (1978). Simple Statistical Models for Discrete Panel Data Developed and Applied to Test the Hypothesis of True State Dependence against the Hypothesis of Spurious State Dependence. *Annales de l'Insee*, nos. 30-31, pp. 227-70.

Heckman, J. (1981a). Statistical Models for Discrete Panel Data, in C. Manski and D. McFadden (eds.), *The Structural Analysis of Discrete Data*. Cambridge, MA: MIT Press, 114-178.

Heckman J. (1981b). The Incidental Parameters Problem and the Problem of Initial Conditions in Estimating a Discrete Time-Discrete Data Stochastic Process in C. Manski and D. McFadden (eds.), *The Structural Analysis of Discrete Data*. Cambridge, MA: MIT Press, 179-198.

Huber, J.C. (1998). Cumulative advantage and success-breeds-success: The value of time pattern analysis. *Journal of the American Society of Information Science* 49: 471-476.

Johnson, N. L. and S. Kotz (1969). *Discrete Distributions*. New York: John Wiley and Sons.

Lotka, A. J. (1926), The Frequency Distribution of Scientific Productivity," *Journal of the Washington Academy of Sciences* 16(12): 317–23.

Macurdy, T. E. (1982). The use of time series processes to model the error structure of earnings in a longitudinal data analysis. *Journal of Econometrics* 18(1): 83-114.

Mairesse, J. and M. Pezzoni (2015). Does gender affect scientific productivity?: A critical review of the empirical evidence and a panel data econometric analysis for French physicists. *Revue Economique* 66: 65-113.

Mairesse, J. M. Pezzoni, and F. Visentin (2019). Does gender matter for promotion in academia?: Evidence from physicists in France. ENSAE-CREST and U Bordeaux: Working paper.

Mairesse, J. and J. Pouget (2008), Citation Hierarchies among Scientific Journals: A Look at Inequality, Persistence and Cumulative Advantage in Impact Factors of Scientific Journals across Disciplines (1981-1994). Presentation to an EPFL conference on science and scientists (September 2008).

Merton R.K. (1988), The Matthew Effect in Science, II: Cumulative Advantage and the Symbolism of Intellectual Property, *Isis* 79(4): 606-623.

Merton R.K. (1968), The Matthew Effect in Science, *Science* 159 (3010): 56-63

Pouget, J. and V. Loonis (2001), Mémoire d'étude en économétrie des panels, Dir. J Mairesse, ENSAE mimeo.

Price, D. de S. (1976), A General Theory of Bibliometric and Other Cumulative Advantages Processes, *Journal of the American Society for Information Science* 27(5): 292-306.

Priem, J., H. Piwowar, and R. Orr (2022). OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts. ArXiv. https://arxiv.org/abs/2205.01833

Scherer, F. M. (1998). The Size Distribution of Profits from Innovation. *Annales d'Economie et de Statistique* 49/50.: 495-516.

Stephan, P. (1996), The Economics of Science, *Journal of Economic Literature* 34: 1199-1235.

Swamy, P. A. V. B. (1970). Efficient inference in a random coefficient regression model. *Econometrica* 38: 311-323.

A. Constructing the sample

The source of our sample was two panel datasets supplied by the CNRS at two different points in time, 2005 and 2015. Each contained the full publishing career to date of the physicists present in that year, as well as their gender, birth year, and rank (Charge de Recherche or Directeur de Recherche, roughly corresponding to assistant and full professor ranks at the university). These data have already been analyzed in other ways by Mairesse and Pezzoni (2015) and Mairesse et al. (2019).

Although we started with these physicists as our sample, the actual data on number of papers published each year and number of citations was newly obtained from OpenAlex (December 2022, February 2023, and February 2024 editions, Priem et al. 2022). This means that the time series for each physicist is no longer censored by failure to appear in one of the panels.

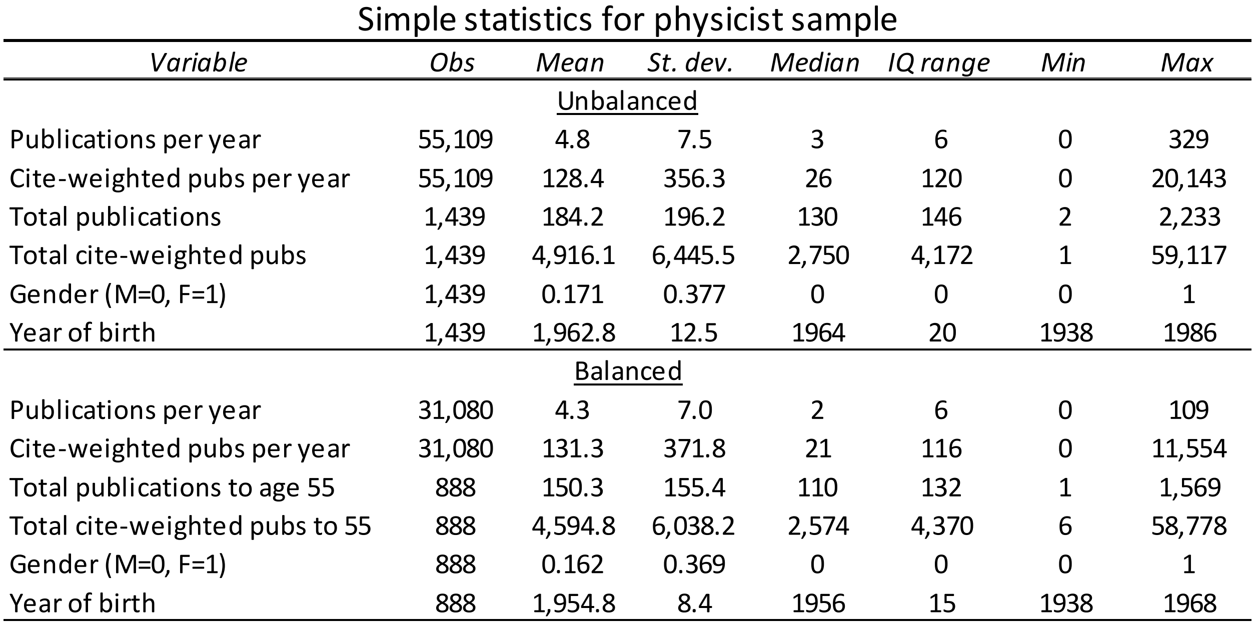
The complete population of CNRS physicists numbered 1,490. We lost a few physicists either because name disambiguation was impossible (for example, Yuan Lu and Christopher Smith) or because we found no data on Open Alex.[10] Our population was therefore 1,479 physicists, approximately 17 per cent of whom are female. To construct our sample, we required that their first publication be age 45 or earlier and that they have at least one cite to their publications over the whole period.[11] This reduced our estimation sample to 1,439 physicists, or 95 per cent of the population. A probit for the probability of being in the sample as a function of birth year and gender revealed that there was no selectivity with respect to those characteristics (which are the only ones we have for physicists that were not found on Open Alex).

We use two versions of the sample data: an unbalanced version that includes as many observations as possible for each physicist (up to 64 years) and a balanced version that is truncated at 35 years of data (age 21 to 55) for each physicist. We show some simple statistics for these two samples below. The top panel includes all years for each physicist, while the bottom panel truncates each one at age 55. The truncated sample has slightly lower publications per year and slightly higher cite-weighted publications per year, whereas the total number of citations and publications is lower for both, because they are measured at age 55 rather than at the end of the career.

[10] We spent a considerable amount of time searching on the internet for those we could not find on Open Alex, checking for name variants, use of married names, and so forth. This procedure allowed us to retrieve the majority of those not found on a first pass.

[11] It is likely that the majority of those with a first publication after age 45 are missing some of their data, as they frequently had names that were more difficult to disambiguate via web searches.

Table A1



Simple statistics for physicist sample

Variable	*Obs*	*Mean*	*St. dev.*	*Median*	*IQ range*	*Min*	*Max*
Unbalanced							
Publications per year	55,109	4.8	7.5	3	6	0	329
Cite-weighted pubs per year	55,109	128.4	356.3	26	120	0	20,143
Total publications	1,439	184.2	196.2	130	146	2	2,233
Total cite-weighted pubs	1,439	4,916.1	6,445.5	2,750	4,172	1	59,117
Gender (M=0, F=1)	1,439	0.171	0.377	0	0	0	1
Year of birth	1,439	1,962.8	12.5	1964	20	1938	1986
Balanced							
Publications per year	31,080	4.3	7.0	2	6	0	109
Cite-weighted pubs per year	31,080	131.3	371.8	21	116	0	11,554
Total publications to age 55	888	150.3	155.4	110	132	1	1,569
Total cite-weighted pubs to 55	888	4,594.8	6,038.2	2,574	4,370	6	58,778
Gender (M=0, F=1)	888	0.162	0.369	0	0	0	1
Year of birth	888	1,954.8	8.4	1956	15	1938	1968

The reason the balanced panel is so much smaller than the unbalanced panel is that many physicists are too young for us to observed their full productivity until age 55. Note that the birth year for the unbalanced panel ranges from 1938 to 1986, whereas for the balanced panel it stops at 1968 (=2023-55). Correspondingly, the balanced panel has slightly fewer females due to the earlier cohort age and the fact that there is increasing participation by women over time.

Figures A1-A4 show the characteristics of the sample and the balanced sample. Figure A1 shows that the physicists in the sample were born between 1938 and 1986, which means that the youngest in our sample are aged 36 at the end. As discussed above, those in the balanced sample were born n 1968 or earlier. Figure A2 shows that the age of first publication is generally in the mid to late twenties, although there are a few between 30 and 40.

Figure A1

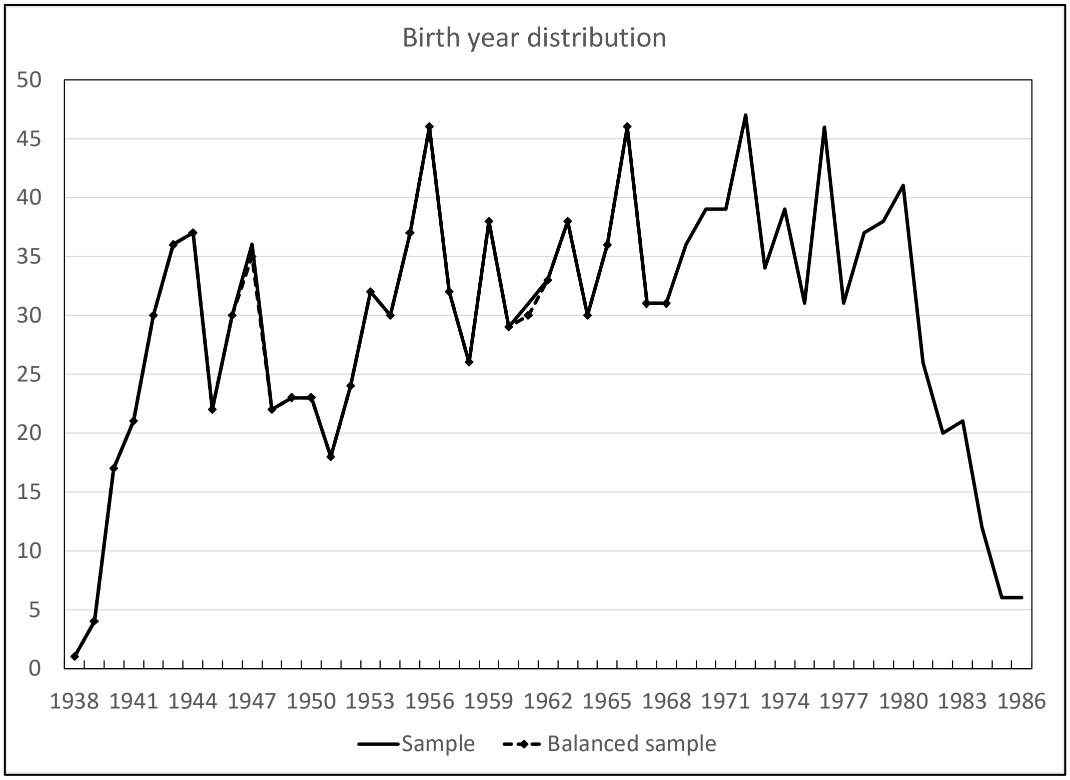


Figure A2

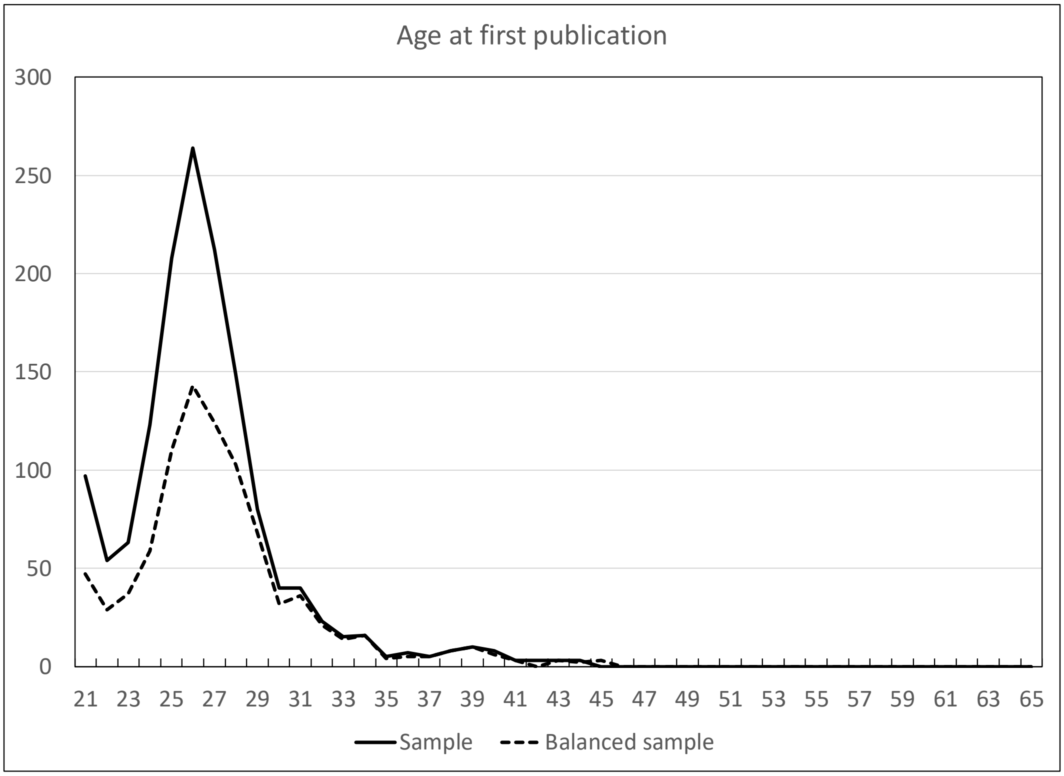


Figure A3 shows that the vast majority of physicists in our sample published their latest paper between 2020 and 2022, although a very few stopped publishing before that. Finally, Figure A4 shows that the latest age at which they publish is more than 40 for the sample, with a significant drop off around age 68 probably due to retirement. Truncation accounts for the large number whose last publication is before then; for the balanced sample the physicists whose last publication is before age 55 have been deleted.

Figure A3

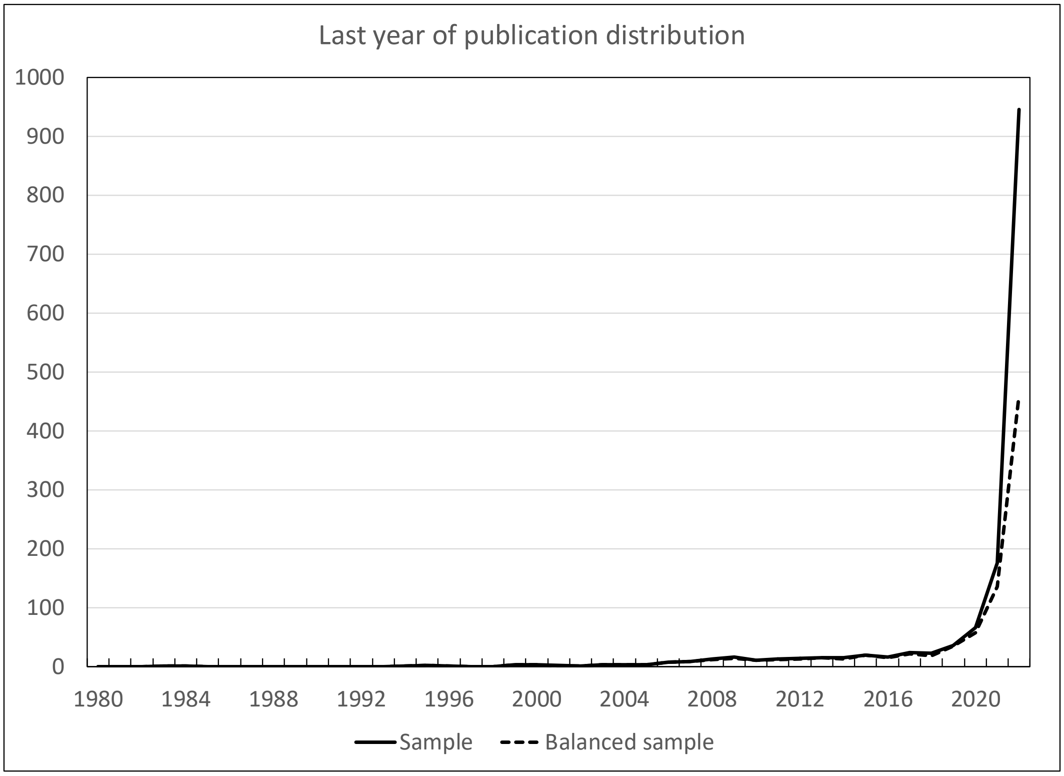
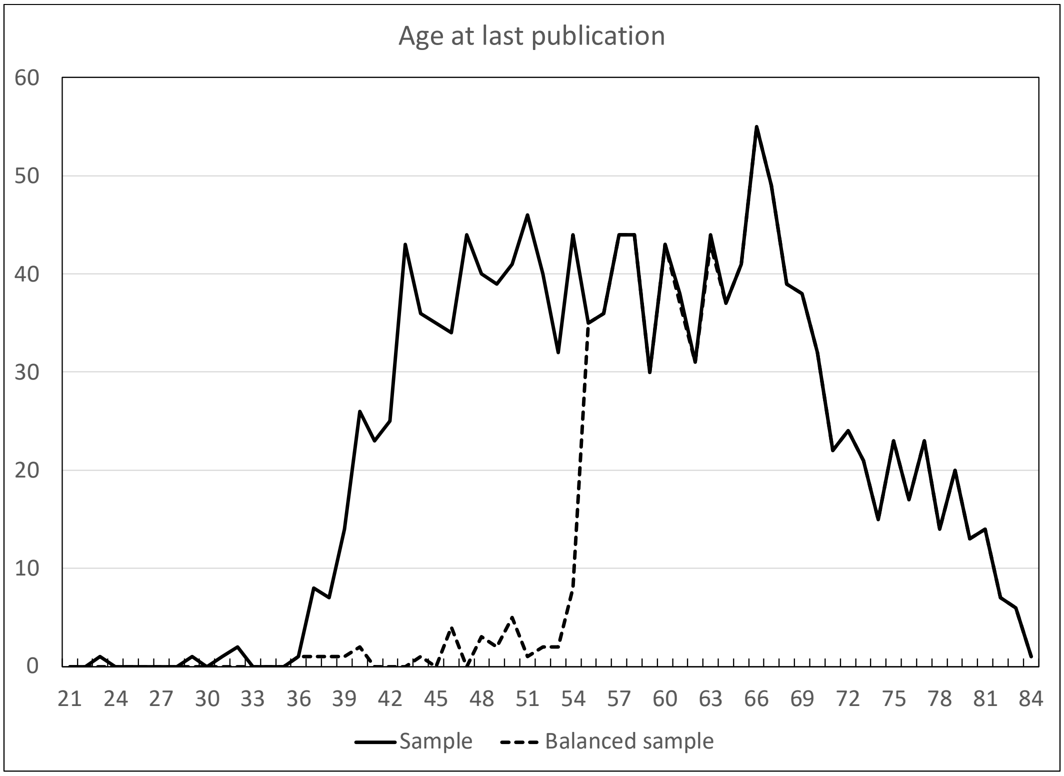


Figure A4



B. Results from estimation on individual physicists' data

Huber (1998) argues that individual inventor output (patents) is fit well by Poisson and does not exhibit overdispersion, but that the individual means differ which produces overdispersion in the aggregate. That is, the data is well-described by initial productivity differences, and observed patents that are determined by a simple Poisson with varying mean productivity. However, his data may have low power at the individual level. We have a much larger and longer sample, so we repeat this analysis here to see if his conclusions still hold.

We estimated a negative binomial model on each physicist to see what the distribution of overdispersion was, and whether the Poisson could be accepted at an individual level. Because the typical publication pattern for a physicist is hump-shaped, we included a quadratic in research age in the negative binomial regression and then tested for overdispersion. We also estimated a gamma model for the individual means, which would be appropriate if the distribution for the panel was negative binomial.[12] Table B-1 presents

[12] The negative binomial distribution can be generated by compounding individual Poissons with means λ_i, $i=1,\ldots,N$ with a gamma distribution on the λs.

some results for three different sets of estimates: the unbalanced and balanced samples with controls for age and age-squared, and the balanced sample without those controls.

Table B-1

Comparison of individual models

	(1)	(2)	(3)
Observations (physicists)	55,109 (1,439)	29,890 (854)	29,890 (854)
Publications			
Share failed t-test for Poisson	52.8%	50.2%	19.7%
Share failed chi-squared for Poisson	50.9%	49.4%	88.8%
Chi-squared for neg. binomial on physicist productivity (df=50)	84.9	59.5	
p-value for neg. binomial on physicist productivity	0.001	0.168	
Cite-weighted publications			
Share failed t-test for Poisson	66.4%	72.0%	97.8%
Share failed chi-squared for Poisson	99.9%	100.0%	100.0%
Chi-squared for neg. binomial on physicist productivity (df=50)	130.1	78.5	
p-value for neg. binomial on physicist productivity	0.000	0.006	

(1) Unbalanced sample for 9-64 years each; Poisson regressions include age and age squared.

(2) Balanced sanple of 35 years each; Poisson regressions include age and age squared.

(3) Balanced sanple of 35 years each; Poisson regressions include only a constant.

The top panel of Table B-1 shows results for the number of publications and the bottom panel for cite-weighted publications. Column (1) contains results for the unbalanced panel, column (2) for the balanced panel with age and age-squared included and column (3) for the balanced panel without controls for age. The first row in each panel shows the share of physicists whose data failed the t-test that the data were Poisson, while the second row shows the share that fails a chi-squared (likelihood ratio) test of the negative binomial versus the Poisson.[13] The bottom two rows in each panel pertain to a binned chi-squared test that the distribution of average productivity across physicists conforms to a gamma distribution. We also display the results from column (1) graphically below, in Figures B-1 and B-2.

[13] In both cases, the critical value used was 5 per cent, so under the null with independent processes for each physicist, we would expect a value here of about 5 per cent.

Figure B-1

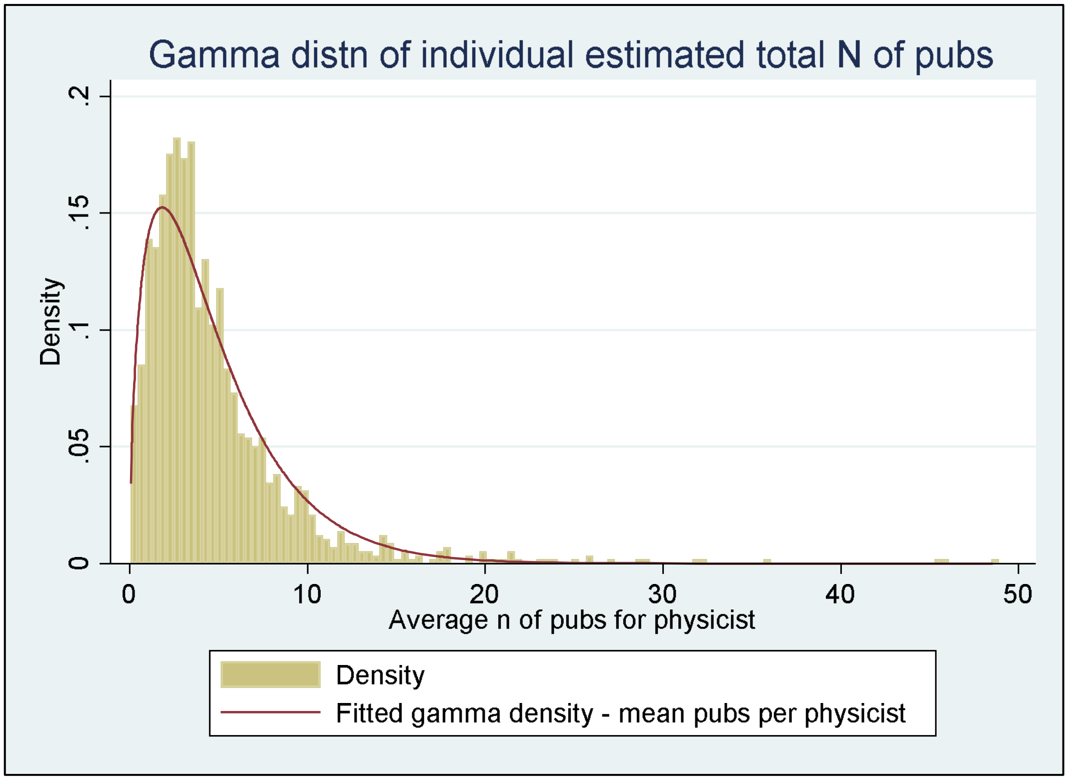
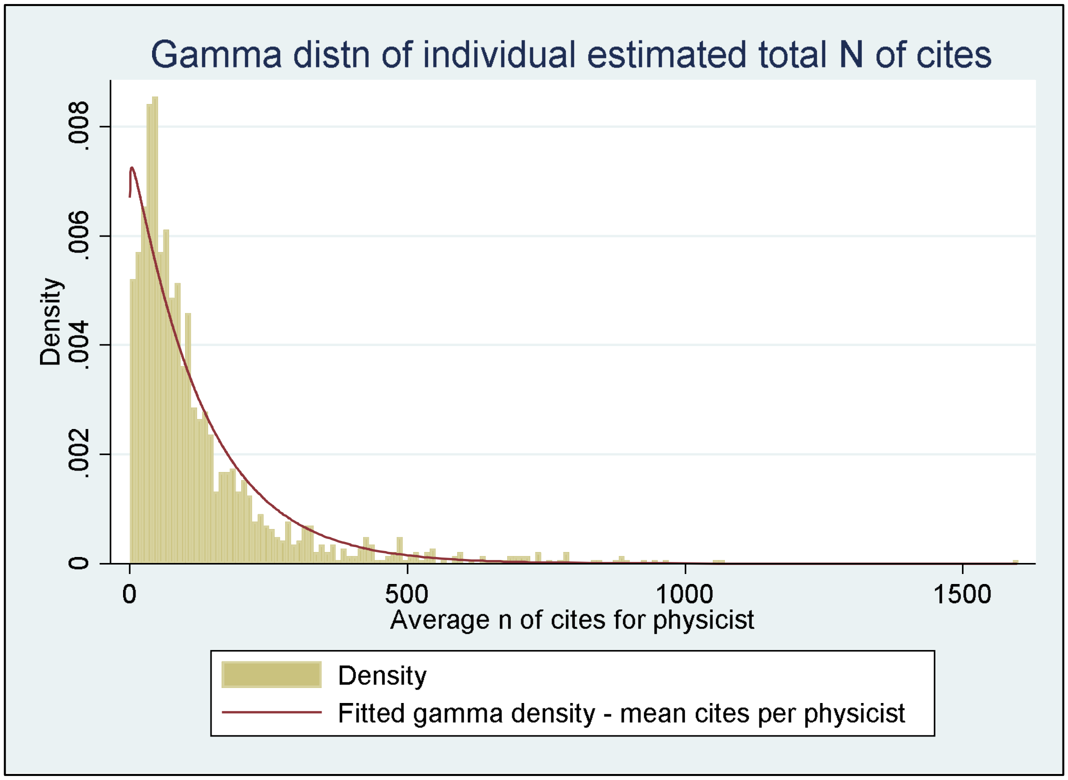


Figure B-2



The form of the log gamma distribution used to estimate the curve is the following:

$$\log f(x;\alpha,\beta) = (\alpha-1)\log x - \beta x + \alpha\log\beta - \log\Gamma(\alpha)$$

The estimated parameters of the gamma distribution for publications are *α = 1.63* and *β = 2.91* with a mean of 4.74 and variance of 18.55. For cite-weighted publications, the parameters were *α = 1.03* and *β = 123.9* with a mean of 127.1 and variance of 15,887.

Our conclusion is that a simple Poisson model does not fit the individual paper count time series, unweighted or weighted by citation counts. Nevertheless, the distribution of average paper counts (both weighted and unweighted) is fairly close to a gamma distribution, especially during the first 30 years of the career (the balanced sample), although gamma is marginally rejected. The implication is that a more complex process than Poisson is driving both the paper and citation count distribution, leaving room for cumulative advantage.

C. Modeling the data parametrically

This appendix describes an exploration of the distribution function for the publication and impact factor data. We examined four simple functions: 1) Poisson; 2) Negative binomial; 3) Log normal; and 4) Pareto. The first and last are one parameter functions and the second and third are two parameters. The densities for these four models are the following:

$$Poisson:\ f(x) = \frac{e^{-\lambda}\lambda^{x}}{x!}$$

$$Negative\ binomial:\ f(x) = \frac{\Gamma(\alpha + x)}{\Gamma(\alpha)\Gamma(x+1)}\left(\frac{1}{1+\beta}\right)^{\alpha}\left(\frac{\beta}{1+\beta}\right)^{x}$$

$$Lognormal:\ f(x) = (x\sqrt{2\pi\sigma})^{-1}\exp\left[-\frac{1}{2}((\log x - \mu)/\sigma)^{2}\right]$$

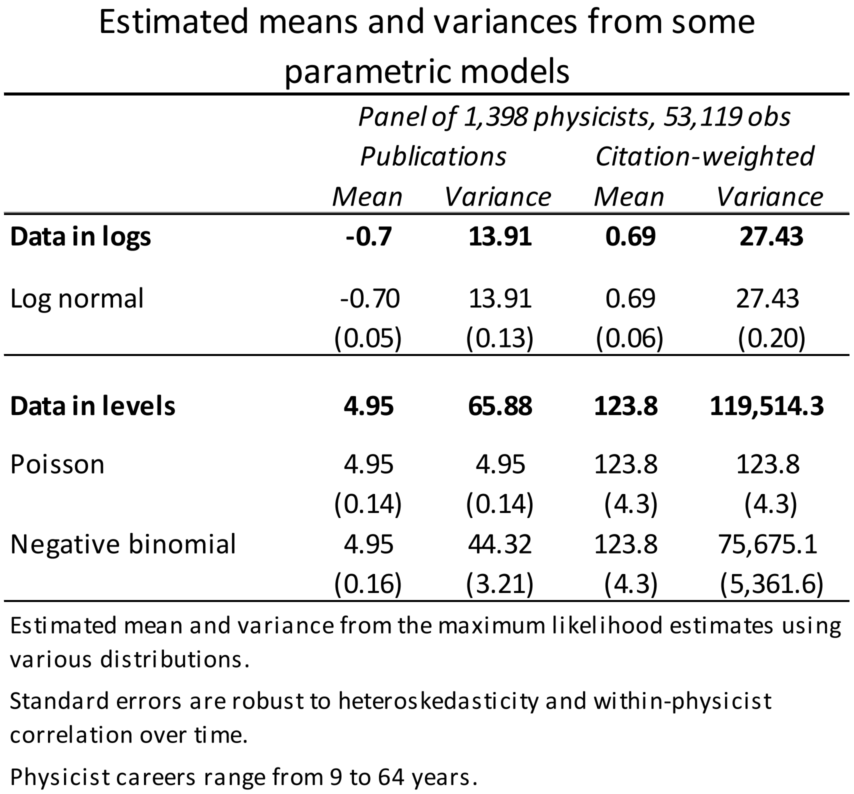
$$Pareto:\ f(x) = \frac{ak^{a}}{x^{a+1}}$$

The parameter *k* for the Pareto is simply the minimum of the distribution, and cannot be estimated using conventional maximum likelihood methods (the likelihood is monotonic in *k*, so the maximum is at the lower bound). The minimum value of *x* in the data is zero, so in the estimation, we set *k* to 0.01.[14]

We estimated each of these models on our two datasets: annual publications and citation-weighted publications for each physicist. Heteroskedastic standard error estimates were clustered by physicist. To compare the models, we computed the means and variances based on the parameter estimates and the results are shown in Table C1. Three things to note about the results: 1) the estimated Pareto coefficient in all cases was significantly less than one, implying no mean or variance, so we do not show those results. 2) the implied means and variances for the level data were very large in the case of the log normal model, so we show the mean and variance for the log data in that case. 3) These simple models can always match the mean of the data exactly; the only deviations from the data are the variances which clearly show overdispersion relative to the simple Poisson or even the negative binomial.

[14] Clearly the likelihood is degenerate for *k=0*, so we use a small number for *k*. The estimates are not very sensitive to the choice of *k*, as long as it is small (less than one).

Table C1



Estimated means and variances from some parametric models

	Panel of 1,398 physicists, 53,119 obs			
	Publications		*Citation-weighted*	
	Mean	*Variance*	*Mean*	*Variance*
Data in logs	**-0.7**	**13.91**	**0.69**	**27.43**
Log normal	-0.70	13.91	0.69	27.43
	(0.05)	(0.13)	(0.06)	(0.20)
Data in levels	**4.95**	**65.88**	**123.8**	**119,514.3**
Poisson	4.95	4.95	123.8	123.8
	(0.14)	(0.14)	(4.3)	(4.3)
Negative binomial	4.95	44.32	123.8	75,675.1
	(0.16)	(3.21)	(4.3)	(5,361.6)

Estimated mean and variance from the maximum likelihood estimates using various distributions.

Standard errors are robust to heteroskedasticity and within-physicist correlation over time.

Physicist careers range from 9 to 64 years.

We illustrate the implied fit of the data using CDFs for the two estimated distributions together with the empirical CDF in Figures C1 and C2. They are similar and show a clear preference for the negative binomial model on all four figures. The negative binomial model fits the data on publications almost exactly, whereas for the citations it overpredicts at lower levels and has lower dispersion than in the empirical data. Our conclusion is that the negative binomial model is a clear starting point for modeling these data, but that as in Appendix B, citation-weighted publications are more dispersed than would be implied by that model.

Figure C1

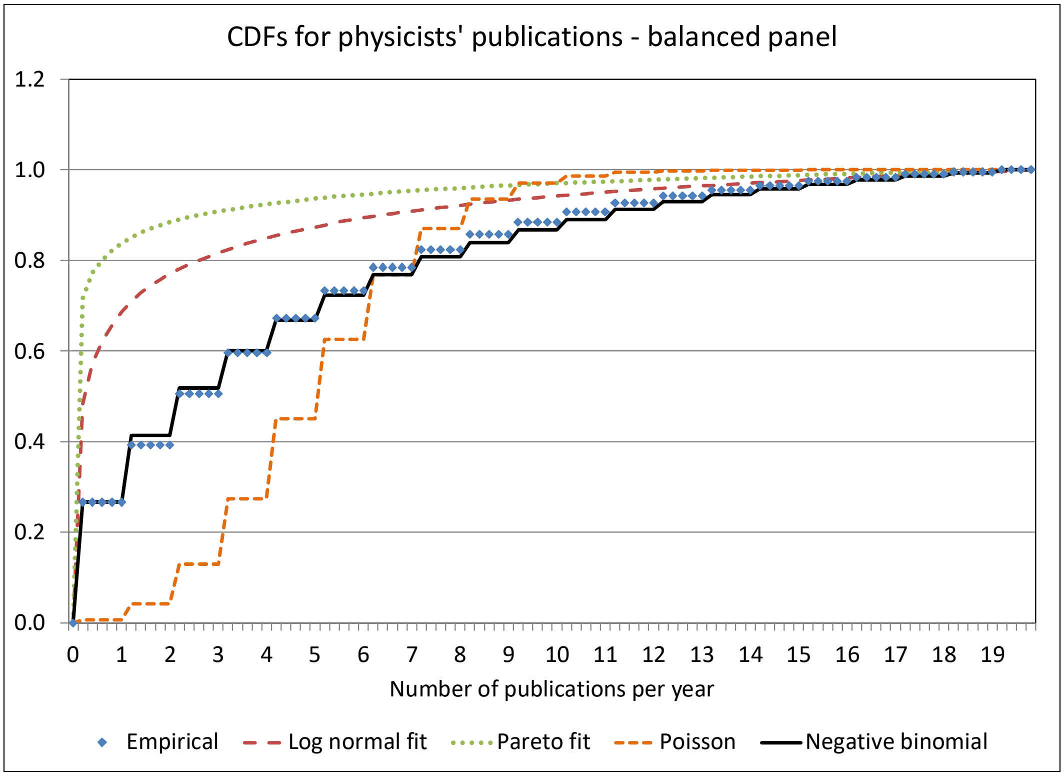
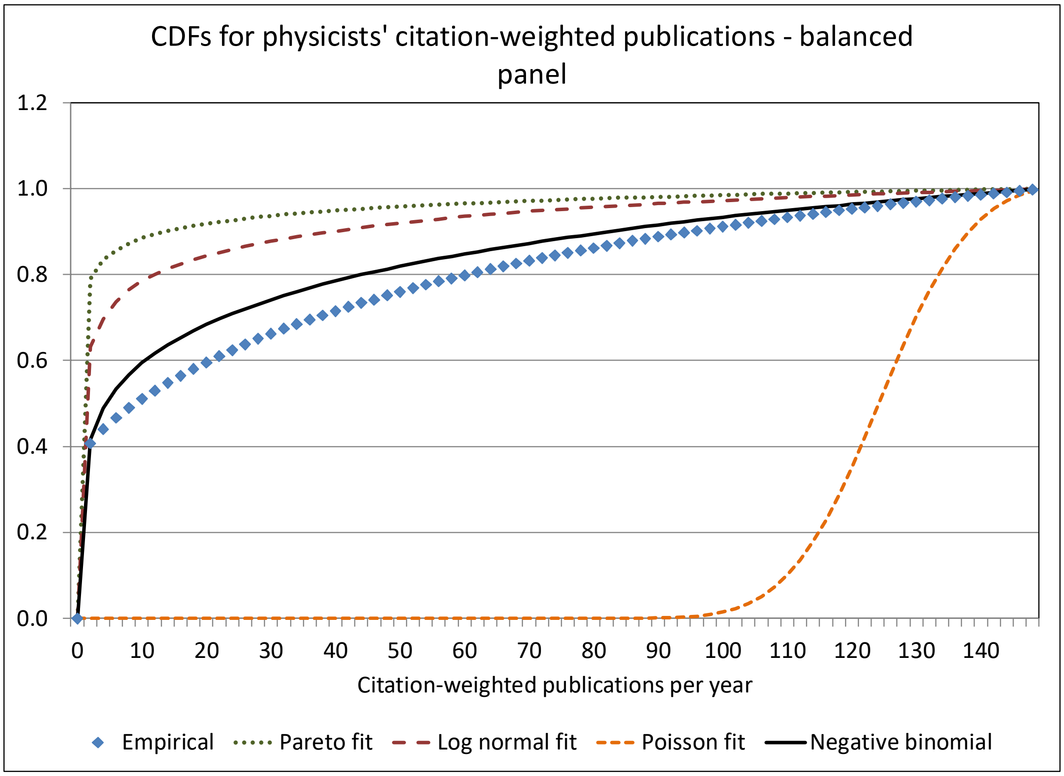


Figure C2



D. Fitting the permanent income model to the journals data

In this appendix we look at these data using the lens of panel time series analysis, using a panel time series model familiar from economics, the permanent-transitory income model. This model is based on a component that evolves as a random walk (the permanent part), which is suitable for capturing the persistence in citations. To this is added a transitory "measurement" error. The model is the following:

$$\begin{aligned} y_{it} &= y_{it}^{p} + u_{it} \qquad\qquad i=1,\dots,N; t=1,\dots,T \\ y_{it}^{p} &= y_{i,t-1}^{p} + v_{it} \end{aligned}$$

where *v* is a permanent shock orthogonal to *u*, and the variance of *u* and *v* for fixed *t* are assumed common among all physicists, but differ across *t*. The physicist is denoted by *i* and the year by *t*, with the number of physicists *N* large and the number of years *T* relatively short. We will therefore try to allow for time-varying variances although it turns out that not all of them are identified in this simple model.

We assume that *u* and *v* are serially uncorrelated. That is,

$$u \sim \left[0, \begin{pmatrix} \sigma_1^2 & 0 & \dots \\ 0 & \sigma_2^2 & \dots \\ \dots & \dots & \dots \end{pmatrix}\right]$$

$$v \sim \left[0, \begin{pmatrix} \varphi_1^2 & 0 & \dots \\ 0 & \varphi_2^2 & \dots \\ \dots & \dots & \dots \end{pmatrix}\right]$$

Differencing, we obtain the following model:

$$\Delta y_{it} = u_{it} - u_{i,t-1} + v_{it}$$

With variance covariance matrix for the journal vector *Δy*:

$$Var(\Delta y) = \begin{bmatrix} \sigma_2^2 + \sigma_1^2 + \varphi_2^2 & -\sigma_1^2 & 0 & \dots \\ -\sigma_1^2 & \sigma_3^2 + \sigma_2^2 + \varphi_3^2 & -\sigma_2^2 & 0 \\ 0 & -\sigma_2^2 & \sigma_4^2 + \sigma_3^2 + \varphi_4^2 & -\sigma_3^2 \\ \dots & 0 & -\sigma_3^2 & \dots \end{bmatrix}$$

We estimate this model by generalized least squares using the second moments as the data observations, and the fourth moments as the weighting matrix. We ignore all the covariances beyond the first one, as they are predicted to be zero and will not contribute to the estimation. In practice, they are zero to three decimal points in the data.

In what follows, we also generalize the model slightly to allow for a coefficient on the "permanent income" (in this case, the permanent paper productivity):

$$\begin{aligned} y_{it} &= \beta y_{it}^{p} + u_{it} \qquad i = 1,\ldots,N; t = 1,\ldots,T \\ y_{it}^{p} &= \beta(y_{i,t-1} - \nu_{i,t-1} + \nu_{it}) + u_{it} \end{aligned}$$

We show in Appendix E that this model is equivalent to an ARMA(1,1) model for *y*. The generalization comes at a cost, because the model must now be estimated in levels, with a far more complex covariance matrix structure, one which is nonzero in all its elements.

1. Identification

Assume that we have *T* periods of data on *y*, and therefore *T-1* periods on *Δy*. This implies *T-1* equations for the variance and *T-2* for the first covariance. There are potentially *T* parameters σ^2 and *T-1* parameters φ^2 to be estimated, which is two too many, regardless of the length of the series. One solution is to impose constancy on the "measurement error" variance, as we do here. But it would also be possible to model the trend in these parameters in order to impose constraints.

2. Results

The variables analyzed are the logarithms of the publications and cite-weighted publications for each physicists and year. We use the unbalanced panel of 1439 observations. First we computed the correlogram for the level and growth of the publications and cite-weighted publications, removing overall year means before doing so:

Figure D-1

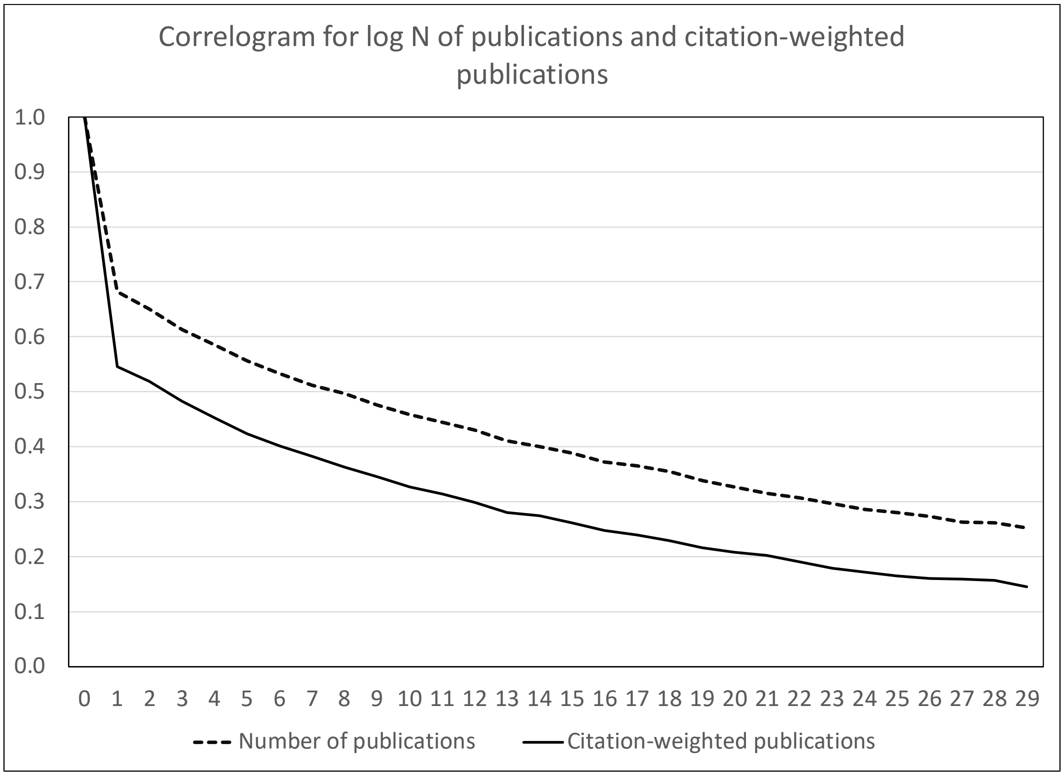
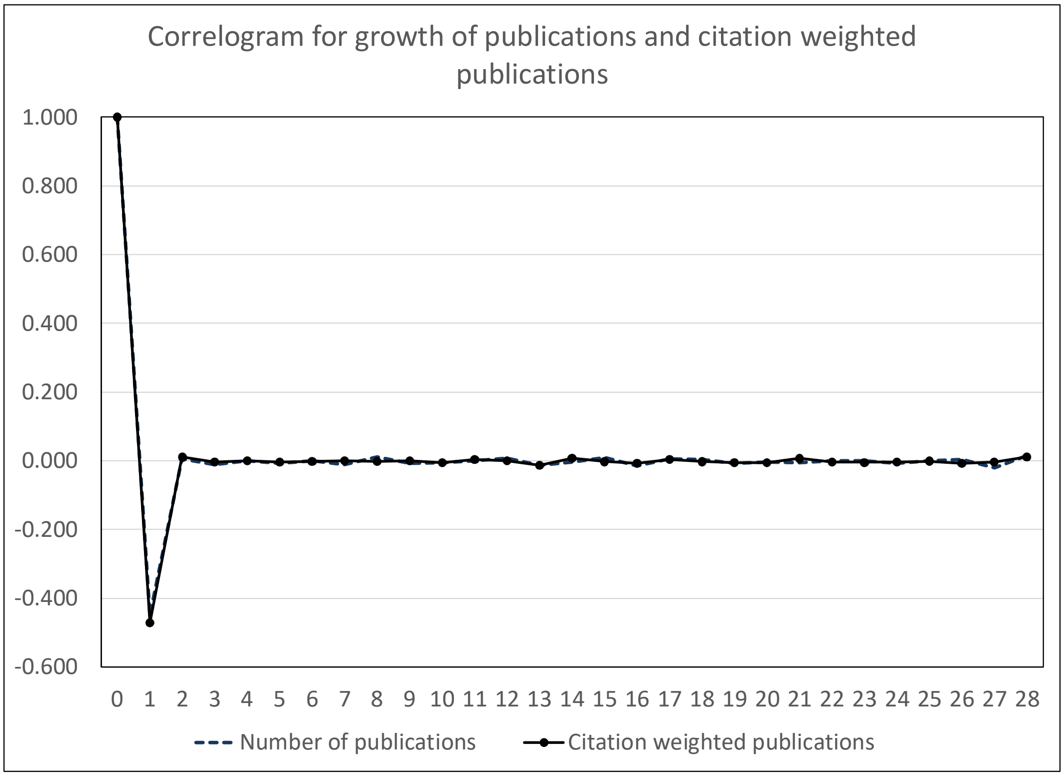


Figure D-2



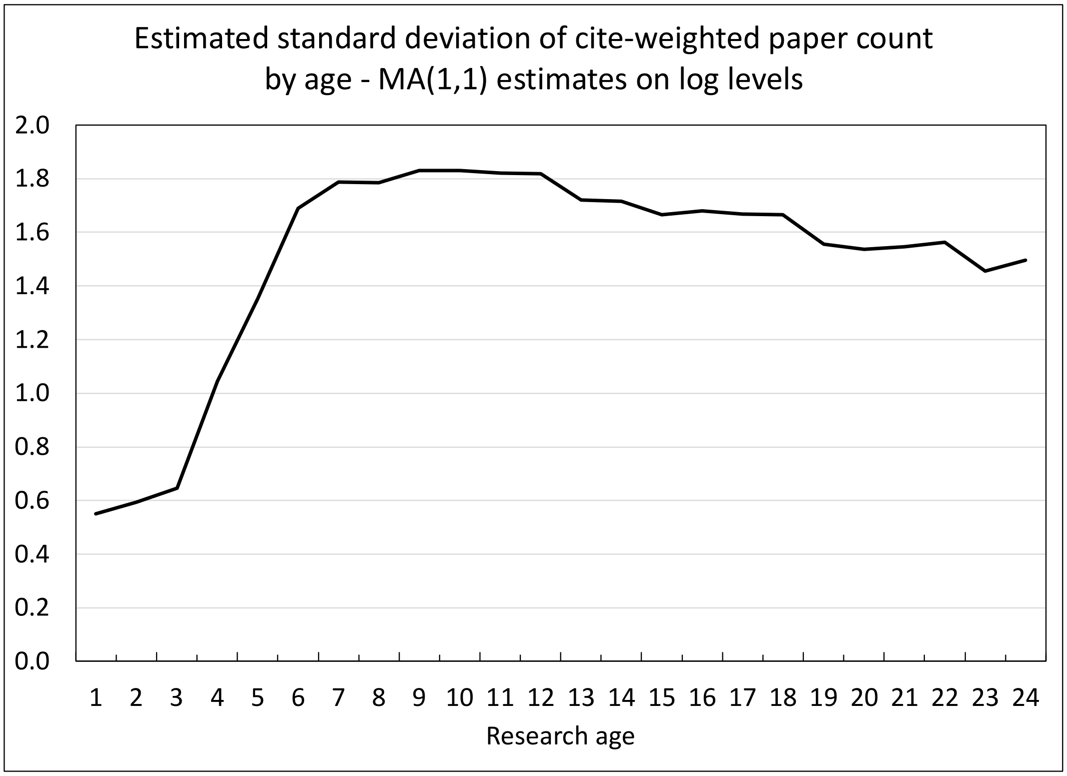
These correlograms clearly show the structure one would expect given the permanent income model sketched above. We can estimate the ratio of φ^2 to σ^2 for both publications and cite-weighted publications using the following equations and assuming constant variances:

$$Var(\Delta y) = 2\sigma^2 + \varphi^2 \approx 0.470, \quad 3.88$$
$$Cov(\Delta y, \Delta y_{-1}) = -\sigma^2 \approx -0.212, \quad -1.82$$

Thus φ^2 is approximately 0.046 for publications and 0.24 for cite-weighted publications and the ratios of the "true" variance to the measurement error variance are approximately 0.22 and 0.13 respectively.

Using the log of cite-weighted papers, the MA(1) model estimates with small standard errors. The main result is that the variance of the permanent component of publications and cite-weighted publications declines slightly over time after the initial increase, which contradicts the idea of increasing dispersion in performance. Here is a graph of the results:

Figure D-3



3. Time series models for panel data

The permanent-transitory model above is a special case of an ARMA(p,q) model with *p=1* and *q=1* (Hall 1987; Ejrnaes and Browning, 2014). Appendix E derives the relationship between the transitory-permanent and ARMA (1,1) model in more detail and we use the ARMA form here in order to rely on known estimation results for that model. Although there are a number of references from the 1980s describing how to estimate panel ARMA models, the key reference for implementation is Macurdy (1982). The basic idea is to form a covariance matrix of the relevant variable(s), including all covariances for a given individual (firm or journal). One then expresses the elements of the covariance matrix as functions of the ARMA parameters and the variance of the disturbance. Estimation is performed by GLS on the covariance matrix with the fourth moments of the matrix as weights.

Due to the N-asymptotics, one can allow variances and other parameters to vary over time. However, because there are N separate time series, it is also necessary to assume a distribution of initial conditions, as otherwise they create an incidental parameters problem. Macurdy shows that if one assumes *p* (the order of the AR) initial conditions distributed with mean zero and variance r_k, $k=1,\ldots,p$, it is possible to estimate a consistent version of the model. The idea is not that new, but it is useful to have the result in this context. In our case, this means there is a single ($p=1$) free variance for the initial condition.

This model was implemented in TSP (Hall and Cummins 2017) using the SUR command, with the results shown in Table D-1 below.

Table D-1

Panel time series estimates for citation-weighted publications

	MA(1) on differences		MA(1) on levels		ARMA(1,1)	
	Coeff.	S.E.	Coeff.	S.E.	Coeff.	S.E.
alpha	1.000		1.000		0.995	0.003
mu	0.580	0.017	0.816	0.006	0.798	0.014
Variance of initial condition	NA		3.910	0.320	3.716	0.346
SIG1	0.500	0.024	0.551	0.042	0.549	0.043
SIG2	0.237	0.018	0.590	0.033	0.592	0.033
SIG3	0.356	0.009	0.639	0.031	0.644	0.031
SIG4	0.005	1.157	1.037	0.032	1.044	0.032
SIG5	0.225	0.028	1.348	0.028	1.354	0.028
SIG6	-0.293	0.024	1.685	0.028	1.689	0.029
SIG7	0.495	0.016	1.784	0.030	1.787	0.030
SIG8	0.438	0.019	1.782	0.031	1.785	0.031
SIG9	0.689	0.014	1.830	0.031	1.831	0.031
SIG10	0.487	0.018	1.830	0.032	1.830	0.032
SIG11	-0.248	0.034	1.823	0.032	1.821	0.032
SIG12	0.283	0.033	1.822	0.032	1.818	0.033
SIG13	0.305	0.030	1.722	0.031	1.721	0.031
SIG14	0.291	0.030	1.721	0.030	1.716	0.030
SIG15	0.391	0.022	1.671	0.033	1.667	0.033
SIG16	0.550	0.015	1.682	0.031	1.679	0.031
SIG17	0.072	0.108	1.668	0.033	1.667	0.033
SIG18	0.401	0.022	1.665	0.031	1.665	0.032
SIG19	0.143	0.057	1.558	0.033	1.556	0.033
SIG20	0.672	0.013	1.536	0.033	1.537	0.033
SIG21	0.299	0.026	1.547	0.034	1.547	0.034
SIG22	0.403	0.022	1.560	0.032	1.562	0.032
SIG23	0.318	0.019	1.453	0.033	1.456	0.033
SIG24	0.318	0.019	1.504	0.030	1.496	0.030
ratio*	0.233	0.019	0.040	0.003	0.047	0.006
Chisq (df)	11001.5 (21)		278.0 (84)		276.0 (83)	
p-value	0.000		0.000		0.000	

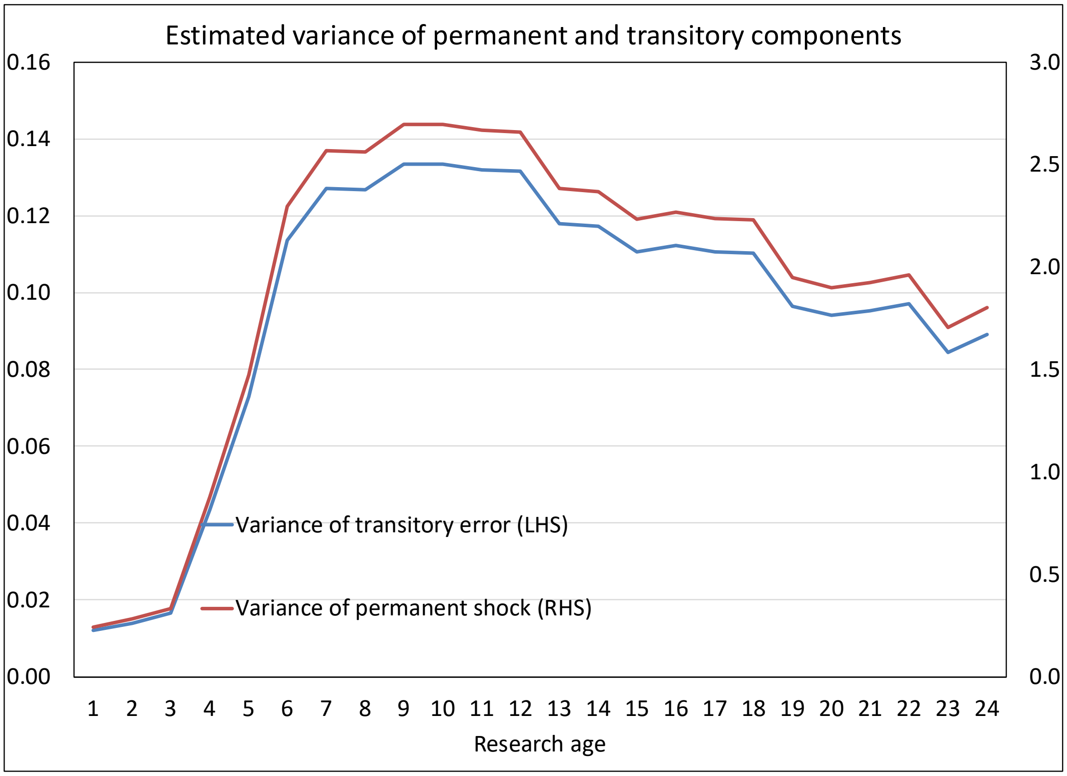
* ratio of signal to total variance (derived).

Sample is 1287 observations on 24 years

Method of estimation is GLS on second moments.

The estimated permanent and transitory variances from the ARMA(1,1) model are shown in Figure D-4 below.

Figure D-4



The conclusion is that the model generates rapidly increasing variance across the physicists in the first seven years, followed by a very slow almost stable variance in the shocks afterwards.

E. Time series models for panel data

In this appendix, we show that the permanent-transitory income model is observationally equivalent to an ARMA(1,1) model. The details of the Macurdy (1982) method of estimating this type of model using panel data are then presented.

Suppressing the physicist (i) subscript for the moment, the permanent-transitory (P-T) income model has the following general form:

$$y(t) = y^p(t) + \nu(t) \qquad t = 1,\ldots,T$$
$$y^p(t) = \beta y^p(t-1) + u(t) \qquad \beta \equiv 1$$

where *v* is a permanent shock orthogonal to *u* and y^p is unobservable permanent income. Eliminating y^p, this model can be rewritten as follows :

$$y(t)-\beta y(t-1)=v(t)-\beta v(t-1)+u(t)$$

The ARMA(1,1) model has this slightly different form:

$$y(t)-\alpha y(t-1)=\varepsilon(t)-\mu\varepsilon(t-1)$$

If *u, v,* and *ε* are normal, these two processes are equivalent. Even if the distributions are non-normal, the first two moments coincide, as both the left-hand and right-hand sides of the two above equations are equivalent. It is possible to show this rigorously using the full covariance matrix for *y* and to derive the identities that relate the unknown parameters in the two processes:

$$\begin{aligned}\beta &= \alpha \\ \sigma_v^2 &= (\mu/\alpha)\sigma_\varepsilon^2 \\ \sigma_u^2 &= (1+\mu^2)\sigma_\varepsilon^2-(1+\alpha^2)\sigma_v^2 \\ &= \left[(1+\mu^2)-(1+\alpha^2)(\mu/\alpha)\right]\sigma_\varepsilon^2\end{aligned}$$

For variances to be positive, these equations impose a condition on the parameters of the ARMA(1,1) process: the sign of *μ* and *α* must be the same (the second equation).

Macurdy (1982) shows how to estimate this model using panel data with large N (the number of units in the panel) and finite T (the number of time periods). To implement his method for our data, I use equations (12) and (13) from his paper with $v_{it} = y_{it}\text{-}y_{.t}$ (year means already removed from the dependent variable *y*), *p = q = 1, T=24,* and a simple change of notation from *a* to *α* and *m* to *μ.*[15] Without loss of generality, I also normalize μ_0 at unity so there is only a single moving average parameter $\mu=\mu_1$. Equation (12) becomes the following, where I have suppressed the stacked notation over time:

$$v_i(t)=\alpha v_i(t-1)+\varepsilon_i(t)-\mu\varepsilon_i(t-1) \qquad t=1,\ldots,T \quad i=1,\ldots,N$$

In equation (13), Macurdy shows via repeated substitution for lagged *v* that this model is equivalent to the following:

[15] Note that I have reversed the signs of mu and alpha from his presentation, in order to conform to conventional ARMA(1,1) notation.

$$\begin{aligned} & v_i(t) = \sum_{j=0}^{t-1} B_j \varepsilon_j (t-j) + B_t r_i(0) \\ & B_0 = 1 \\ & B_1 = \alpha - \mu \\ & B_j = \alpha B_{j-1} \quad j = 2, \ldots, \mathrm{T} \end{aligned}$$

Because the order of the autoregressive process is one, there is one initial condition for this model, *r_i(0)*. Given a panel of observations and the assumption of a random initial condition, the model has unknown parameters *α, μ,* the variance of the initial condition σ_0^2 , and the variances of *ε,* σ_t^2 *, t=1,...,T.* By constraining *α* to be unity, one can also use the same model to estimate the random walk MA(1) model on level data. This is subtly different from estimating the MA(1) model on differenced data, because the latter does not allow estimation of the variance of the initial condition.

The proposed estimation method is quasi-maximum likelihood, which is a version of method of moments. In the case where the variances of the random draw ε are allowed to vary over time, we form each element of the covariance matrix for each year t separately. For our ARMA(1,1) model, the equation above yields the following sequence for *v*:

$$\begin{aligned} & v_i(1) = B_0 \varepsilon_i(1) + B_1 r_i(0) \\ & v_i(2) = B_0 \varepsilon_i(2) + B_1 \varepsilon_i(1) + B_2 r_i(0) \\ & v_i(3) = B_0 \varepsilon_i(3) + B_1 \varepsilon_i(2) + B_2 \varepsilon_i(1) + B_3 r_i(0) \\ & \bullet \\ & \bullet \\ & \bullet \end{aligned}$$

The covariance matrix elements are therefore the following:

$$\begin{aligned} & v_i(1) v_i(1) = B_0^2 \sigma_1^2 + B_1^2 \sigma_0^2 \\ & v_i(2) v_i(2) = B_0^2 \sigma_2^2 + B_1^2 \sigma_1^2 + B_2^2 \sigma_0^2 \\ & \ldots \\ & v_i(1) v_i(2) = B_0 B_1 \sigma_1^2 + B_1 B_2 \sigma_0^2 \\ & v_i(2) v_i(3) = B_0 B_1 \sigma_2^2 + B_1 B_2 \sigma_1^2 + B_2 B_3 \sigma_0^2 \\ & \ldots \\ & v_i(1) v_i(3) = B_0 B_2 \sigma_1^2 + B_1 B_3 \sigma_0^2 \\ & v_i(2) v_i(4) = B_0 B_2 \sigma_2^2 + B_1 B_3 \sigma_1^2 + B_2 B_4 \sigma_0^2 \\ & \ldots \end{aligned}$$

And so forth up to the Tth set of observations. We compute the mean of each of these moments in the data, and then match them to the theory using the fourth moments of the data as a weighting matrix. There are T+3 unknown parameters and T*(T-1)/2 moments, so the model is over-identified if T>4.[16]

[16] In practice, when T is fairly large, the number of moments can be quite large (equal to 276 in our case), and those at long lags do not add a lot of information, so it is possible to estimate using only a subset of moments, as we did here. We chose to use the first 110 moments, up to the covariances at the fourth lag.