\

Title: The Specialization of Scientists

Authors: George J. Borjas,1,2† and Kirk B. Doran3*†

Affiliations:
[1]Harvard Kennedy School, Harvard University, Cambridge, MA 02138, USA.
[2]National Bureau of Economic Research, Cambridge, MA 02138, USA.
[3]Department of Economics, University of Notre Dame, Notre Dame, IN 46556, USA.
†These authors contributed equally to this work.
*To whom correspondence should be addressed. Email: kdoran@nd.edu

Abstract: We use about 2 million papers published by 221,843 individual scientists over eight decades to document that scientists have become more specialized over time. Scientists who began their careers in 1975 contributed to 40% fewer areas of knowledge in the first thirty years of their career than did scientists who began in 1945. We demonstrate two factors that affect scientists' range of contribution. First, scientists narrow their range where their peer groups are large. Second, scientists may expand their range through coauthorship.

Main Text:

It has been widely asserted throughout scientific history that individual scientists are becoming more specialized over time. Albert Einstein wrote in 1932 that:

"The area of scientific investigation has been enormously extended, and theoretical knowledge has become vastly more profound in every department of science. But the assimilative power of the human intellect is and remains strictly limited. Hence it was inevitable that the activity of the individual investigator should be confined to a smaller and smaller section of human knowledge" *(1)*.

Despite the frequent assertion of increased specialization, the available evidence is indirect and anecdotal. Perhaps because of the paucity of evidence, there is disagreement about whether the purported narrowing of focus has led to better science. Einstein himself wrote: "every serious scientific worker is painfully conscious of this involuntary relegation to an ever-narrowing sphere of knowledge, which threatens to deprive the investigator of his broad horizon and degrades him to the level of a mechanic" *(1)*. Max Weber, in contrast, wrote in 1919 that: "A really definitive and good accomplishment is today always a specialized accomplishment." *(2)*.

In short, both the increasing specialization of the individual scientist and its causes and consequences have been difficult to verify and explore empirically. The contributions of individual scientists are typically analyzed by using publication databases. But the most frequently studied databases, such as the ISI Web of Science, do not contain narrowly defined subject classifications of individual papers, restricting the subject classification to broad measures (such as "Mathematics" and "Mathematics, Applied") and assigning these classifications by using information on the journal of publication rather than the content of the actual contribution. These data, therefore, do not directly provide evidence on the increasing specialization of scientists.

We analyze the contributions of individual scientists to determine whether their work has, in fact, become more specialized. We introduce a rarely studied database that contains consistent narrowly defined subject classifications for 2,654,013 individual papers and 551,423 individual authors from 1939 to 2010: the American Mathematical Society's *Mathematical Reviews* database. See the Supporting Online Material for a description of the data *(3)*.

The editors of *Mathematical Reviews* assign each published article to one of 63 primary subfields, such as "Information and Communication, Circuits" and "Harmonic Analysis on Euclidean Spaces." This list of subfields has undergone relatively minor changes over time, allowing us to calculate a consistent measure of specialization. *Mathematical Reviews* covers any field of science in which mathematical proofs and statistics predominate, including theoretical physics (mechanics of deformable solids, quantum theory, etc), theoretical economics and econometrics, theoretical computer science, and mathematical biology. Our database, therefore, consists of the universe of published mathematical knowledge over the last half of the 20th century. Hence, we can precisely measure the degree of specialization of all mathematical and theoretical scientists, permitting us to document how the degree of specialization has changed across subsequent cohorts of scientists, over an individual scientist's career, and in response to various forces.

We use two measures of specialization: the number of fields in which an individual scientist has published over a particular span of his or her career, and the Herfindahl index of the scientist's publications over that time span. The Herfindahl index is a commonly used measure of concentration in economics and is defined by $H_i = \sum_f s_{if}^2$, where s_{if} is the share of publications that scientist i published in field f. The Herfindahl index has a maximum of 1 when a scientist has only published in one field, and a minimum of $1/F$ (where F is the number of potential fields) when the publications are evenly distributed across fields *(4)*.

We calculate the measures of specialization over three distinct spans of a scientist's career: the first 10, the first 20, and the first 30 years. We then calculate the average measure of specialization for each cohort of scientists (e.g., scientists whose first publication was in 1950, or 1951, etc.). Figure 1 plots the trend in the average measures of specialization. Regardless of the length of the span, there is a noticeable increase in the specialization of subsequent cohorts of scientists from 1945 through 1990. The typical scientist in the cohorts that entered scientific activity in the late 1940s published in around 3.8 of the 63 fields base over the next 30 years, while those who entered in the late 1970s published in 2.9 fields.

The increase in specialization could be an artifact of sizable changes in the number of papers published by successive cohorts; if a scientist publishes fewer papers, this mechanically decreases the upper bound in the number of fields he or she could contribute to. Levin and Stephan demonstrate that such vintage effects are not typically important in scientific productivity *(5)*. Further, Figure S1 illustrates the number of papers published by successive cohorts of scientists. There is no large decrease in publications by the later, more specialized cohorts; in fact, there is a small increase. Finally, the Supporting Online Material presents a detailed regression analysis that specifically controls for number of papers and documents an increase in specialization from 1945 to 1990 (see Figure S2).

The increase in specialization lengthened the amount of professional experience required for a scientist to start contributing in an additional field. Figure 2 plots the age profile of the total number of fields in which the average scientist in a particular cohort has published. In the earliest cohort of scientists (who began publishing between 1945 and 1956) it took the average scientist 18 years to reach their third field of publication. Among scientists who first published between 1969 and 1980, it took about 30 years to reach the third field.

Although the data indicate increasing specialization between 1945 and 1990, it is harder to determine why the increase occurred. One possibility is that as the market for exchanging scientific knowledge grew larger, individual scientists have an increased incentive to specialize in narrower ranges of knowledge production. Adam Smith famously noted: "As it is the power of exchanging that gives occasion to the division of labour, so the extent of this division must always be limited by the extent of that power, or, in other words, by the extent of the market" *(6)*. Between 1951 and 1960, 26,012 distinct scientists published at least one paper. Between 1991 and 2000, in contrast, 196,617 scientists published at least one paper. This large increase in the "extent of the market" may encourage increasing specialization.

To determine whether specialization and market size may be linked, we consider how each varies across countries. Scientists in each country tend to compete for jobs in a relatively small set of local research institutions and publish a disproportionately large number of their articles in regional journals, leading to a significant degree of "home country bias" *(7)*. If specialization is more likely to arise when there are many scientists in a specific market, there should be greater specialization in countries with larger scientific communities.

Since 1984, the AMS database provides detailed information on the affiliation of the author at the time of publication. We calculate the average measures of specialization over the first 20 years for scientists who first published between 1984 and 1990 in each of 44 countries (i.e., those countries with at least 50 scientists in their cohort). Figure 3A presents the scatter diagram relating the number of fields to the log size of the scientific community in each country. The data suggests that scientists working in countries with larger communities are more specialized: the slope of the regression line suggests that the quintupling in the number of scientists between the 1950s and the 1990s would reduce the average number of fields by 0.14. The relatively small effect suggests that while the Smith hypothesis is important, there are other factors at work. Among the likely candidates is the increasing burden of knowledge, emphasized by Einstein and studied by Jones *(8)*.

Figure 1 also documents that there has been an attenuation of the increase in specialization in the past two decades. The indices of specialization measured over the first 10 years of a scientist's career stopped declining around 1990, and there has even been a slight reversal since. There has been no decrease in the size of the scientific community in recent years that could explain this emerging trend, nor evidence of a decreasing burden of knowledge. Therefore, we hypothesize a new factor determining specialization: the propensity of scientists to work by themselves or in teams. As Wuchty *et al.* established, there has been a rapid increase in the prevalence of teams in the production of scientific knowledge *(9)*. Figure 4A replicates the result of Wuchty *et al.* in our data, showing a dramatic increase in the number of coauthored papers over the last four decades, particularly among papers written by a team of at least 3 scientists. If scientists are more likely to push into new fields of knowledge in a coauthored paper than in solo work, the rapidly increasing proportion of coauthored papers may explain the recent slowdown (and potential reversal) in specialization.

We test this hypothesis by determining which of a scientist's papers contributed to a field where the scientist had never published previously. In Figure 4B, we plot the probability that a given paper written in the first 10 years of a scientist's career is in a brand new field as a function of the number of authors of that paper. Using either the raw data or a detailed regression analysis (reported in Table S1), we find that the greater the number of authors on a paper, the higher the probability that the paper contributes to a new field. Further, Figure 3B confirms this observation in the cross-country data: the average number of fields is larger in countries where teamwork is more prevalent.

Table S2 reports a multivariate regression that relates the degree of specialization of scientists within a country to both the size of the scientific labor force within that country and to the propensity of scientists within that country to work in teams. The results demonstrate that the larger the market, *ceteris paribus*, the narrower the focus of individual scientists, while the greater the propensity to work in teams, *ceteris paribus*, the broader the focus of individual scientists.

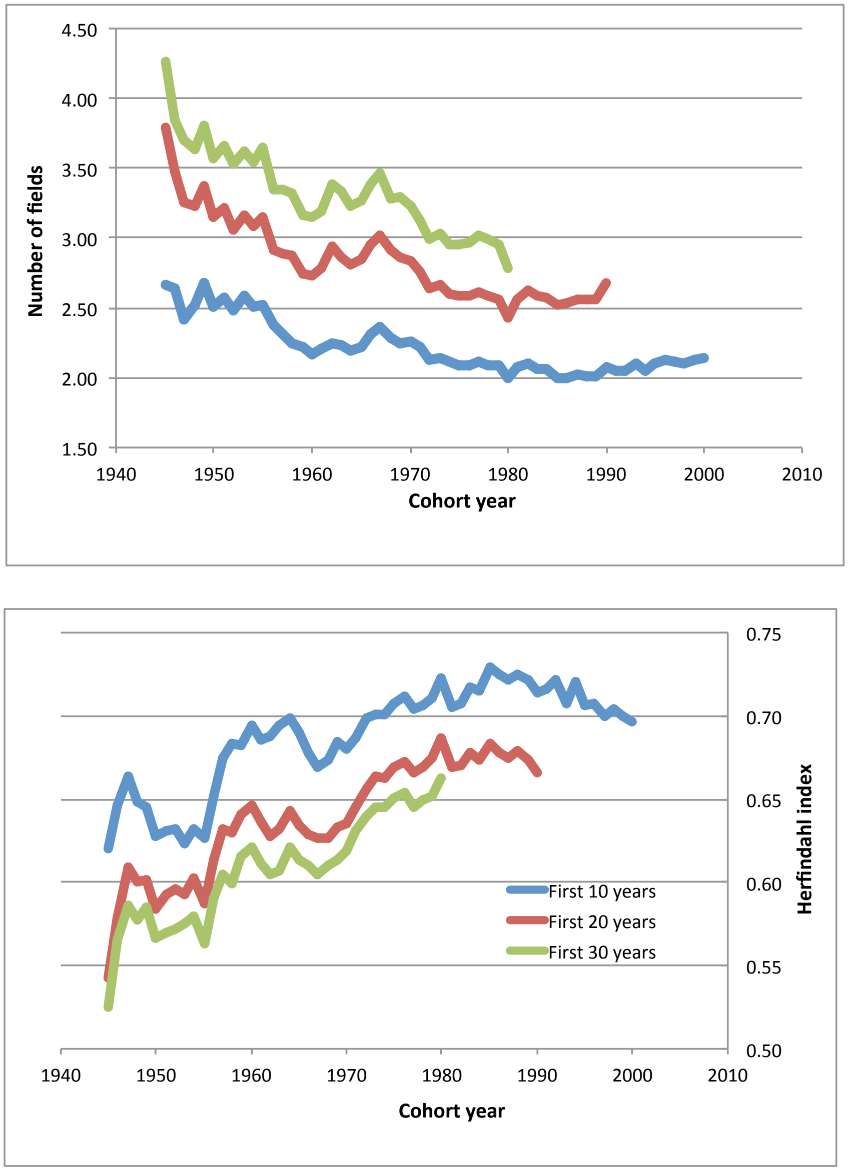
Because field contributions are assigned at the paper-level rather than at the author-level, it is not possible to ascertain if a specific author in a team contributed to the part of the paper that determined its field assignment. For example, a mathematician in a team-authored paper assigned to "fluid mechanics" may have contributed a mathematical proof without actually learning any physics. Thus the relationship between coauthorship and specialization can be interpreted in multiple ways. The regression analysis reported in Table S3 reports that teamwork increases the probability that a scientist enters a new field *and* that the scientist contributes again to that field in future solo-authored work, suggesting that the new fields entered through coauthorship may become part of the authors' permanent expertise. The increasing prevalence of teamwork, therefore, may indeed portend a break in the long-term trend towards ever more specialized work.

References and Notes:

1. A. Einstein, *Ideas and Opinions* (Modern Library, New York, 1994), pp. 74-75.
2. M. Weber, *From Max Weber: Essays in Sociology* (Routledge, New York, ed. 1, 2007), pp. 134-135.
3. We use the subset of the database that consists of papers written by scientists whose first publication was between 1945 and 2000, inclusive.
4. Both measures of concentration are mechanically equal to one if the scientist has only published one paper. Hence, we focus on the subsample of scientists who published at least two articles over the relevant period.
5. Sharon G. Levin, Paula E. Stephan, *The American Economic Review* **81**, 114-132 (1991).
6. A. Smith, *The Wealth of Nations* (Modern Library, New York, 1994), p. 19.
7. George J. Borjas, Kirk B. Doran, The Collapse of the Soviet Union and the Productivity of American Mathematicians. *The Quarterly Journal of Economics*, in press (available at http://www.nber.org/papers/w17800).
8. Benjamin Jones, The Burden of Knowledge and the 'Death of the Renaissance Man': Is Innovation Getting Harder?. *Review of Economic Studies* **76**, 283-317 (2009).
9. S. Wuchty, B. F. Jones, B. Uzzi, *Science* **316**, 1036 (2007).

Acknowledgments: The data used in this paper are available from the American Mathematical Society's MathReviews database (online at: http://www.ams.org/mathscinet/) and Thomson Reuters' ISI Web of Science. This research was supported by grants from the Upjohn Institute, the Kauffman Foundation, and the Alfred P. Sloan Foundation. Both authors contributed equally to this work.

Fig. 1. The increase in specialization. These figures present the average number of fields (A) or the average Herfindahl index (B) over a 10, 20, or 30-year career span for scientists who first published in each calendar year between 1945 and 2000.



A.

B.

Fig. 2. Specialization over the life cycle. These figures present the average number of fields (A) or the average Herfindahl index (B) over the working life for scientists who first published in 1945-56, 1957-68, 1968-80, or 1981-92.

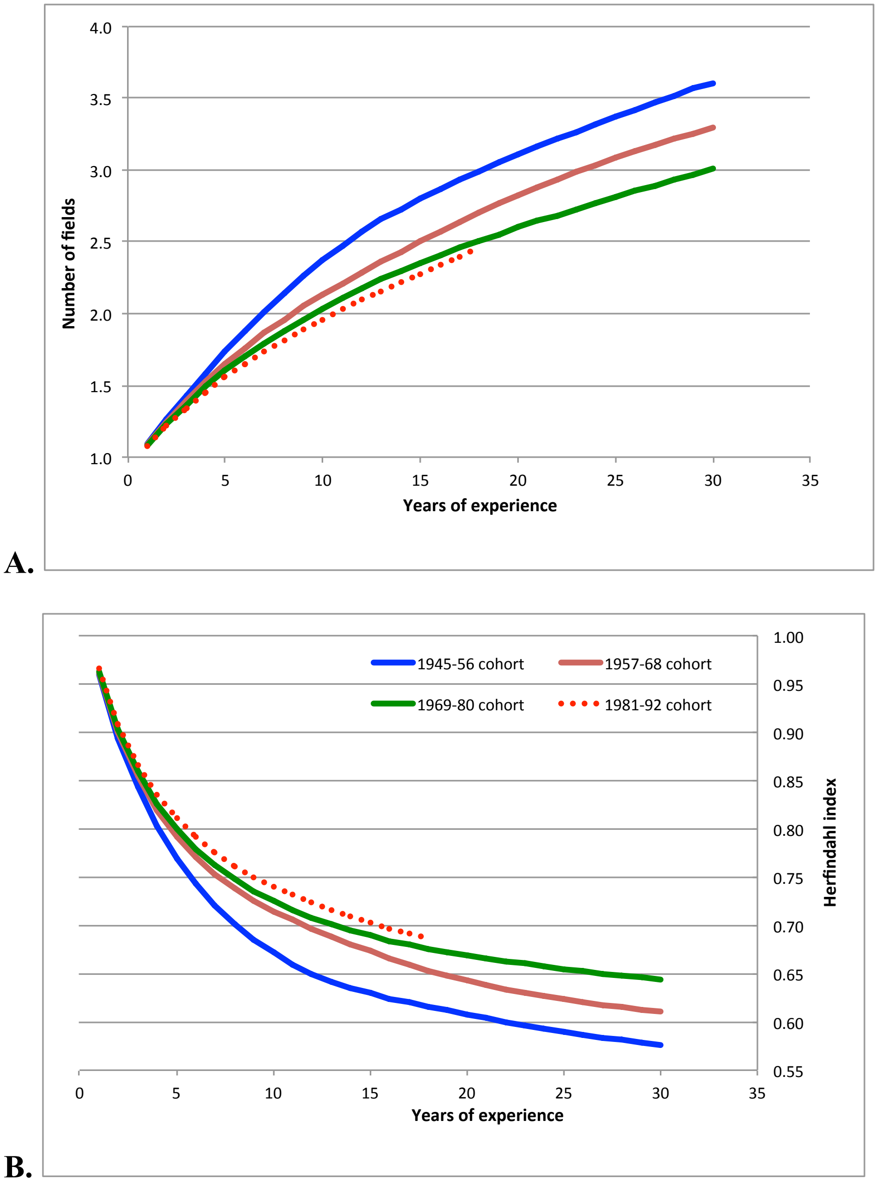


Fig. 3. International differences in specialization. These figures present the relation between the number of fields of scientists in each of 44 countries to the (log) number of active scientists (A) or the fraction of papers in the country written by teams (B). The size of each point in the scatter diagram is proportional to the size of the country's 1984-90 scientific cohort.

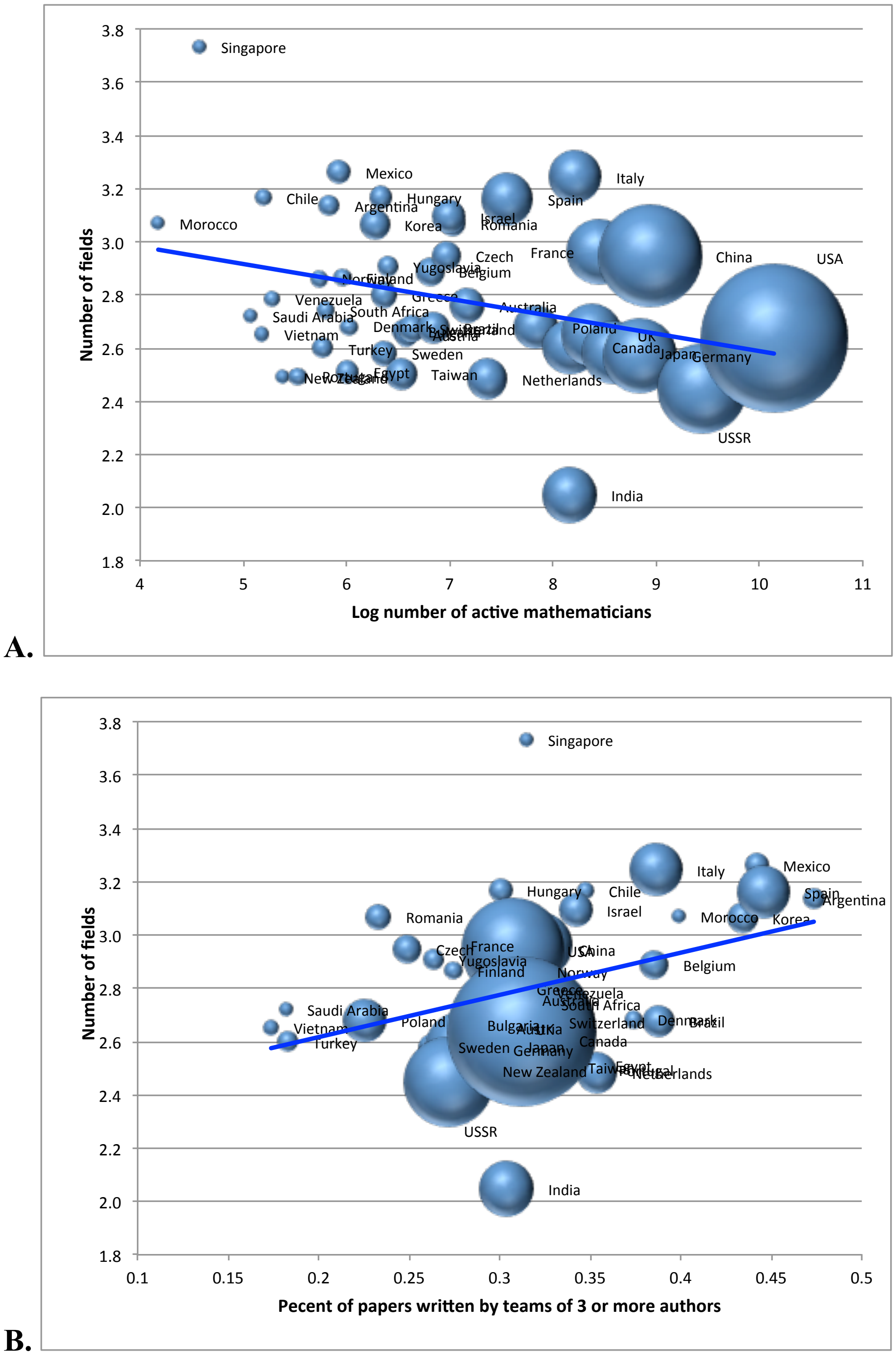
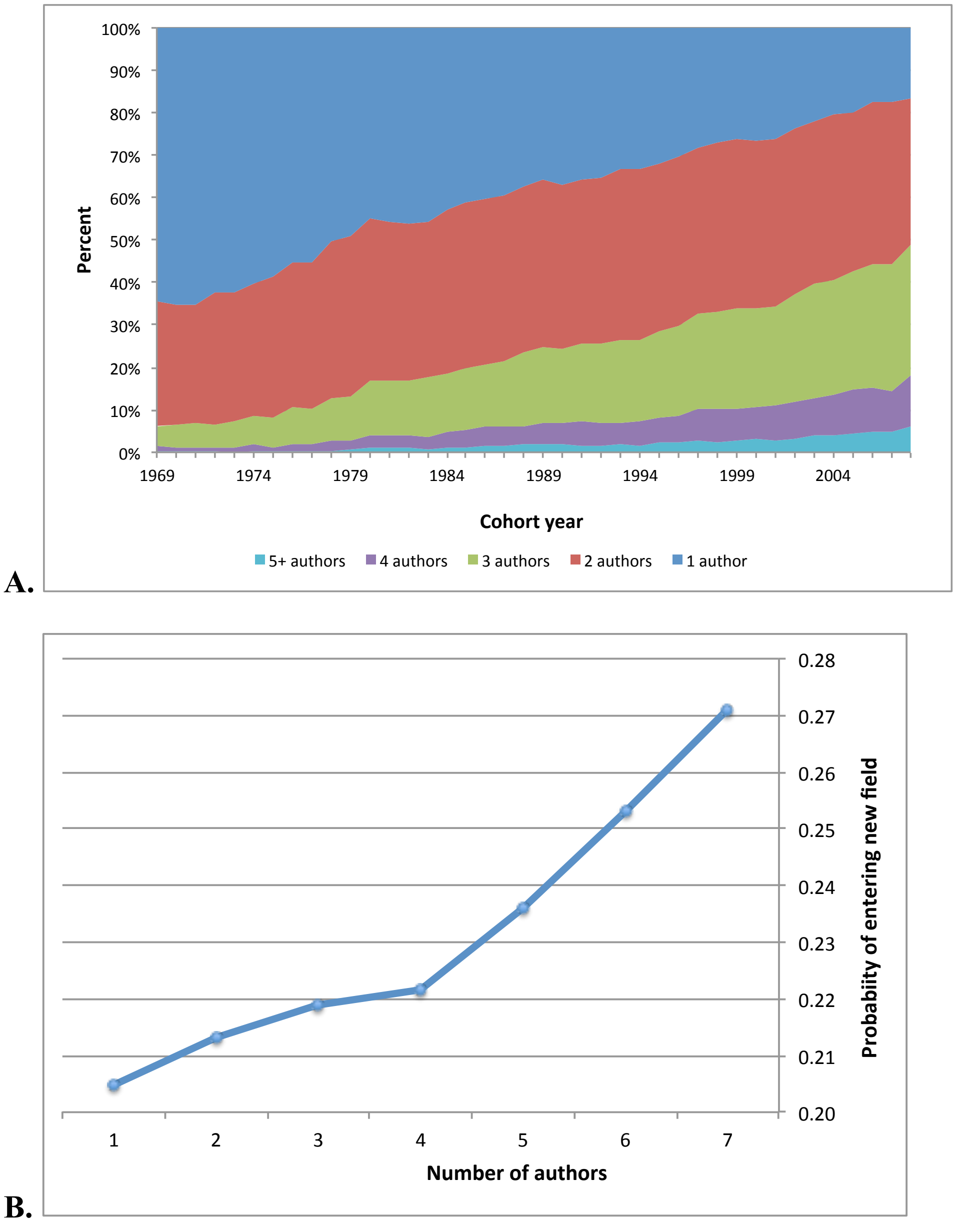


Fig. 4. Specialization and teamwork. These figures present the increase in the fraction of papers that are coauthored (A) and the relation between teamwork and the probability that an additional publication represents an entry into a new field (B).



Supplementary Materials:

Supplementary Materials:

1. Description of data

Our analysis uses two distinct, but related, databases. AMS *Mathematical Reviews* records the titles, publication sources, author names, references, and citations of over 2 million articles from 2,764 different journals and publication sources worldwide. The AMS team assigns the correct set of articles to each person (even in such cases as authors with identical names), tagging each author with a unique author identifier. The AMS also records subject classification codes for each paper in its database: 63 subfields defined according to the 2-digit 2010 Mathematics Subject Classification (MSC2010).

The AMS provided us with a spreadsheet containing all publications in the mathematical and theoretical sciences between 1939 and 2010. A row in this spreadsheet defines a particular permutation of author (i), field (f), and year (t). For each (i, f, t) row we have information on: the number of papers published; the institutional affiliation(s) associated with the papers in that row; and the (country) location of the affiliation(s).

When necessary, we supplement the AMS database with additional information purchased from the Thomson Reuters' ISI Web of Science. The ISI Web of Science records the titles, publication source, author names, references, and citations of millions of articles from thousands of journals worldwide. For many articles (especially after 1978), the database records research addresses and reprint addresses for each author, as well as abstracts, keywords, and funding information.

We purchased the records of all 1,179,787 articles in the primary ISI Web of Science database between 1970 and 2009 for the following categories: Mathematics, Applied Mathematics, Interdisciplinary Applications of Mathematics, Mathematical Physics, and Statistics & Probability. We also purchased all 1,921,587 articles referenced by these main articles, and all the 2,368,123 papers that cite these main articles.

We obtained special permission from the AMS to merge our ISI papers by title, source, and author with the AMS database. We obtained 882,088 matches out of the 1,753,148 journal articles in the AMS database, or slightly over a 50 percent match rate. The merged database allows us to uniquely identify all of the coauthors for each paper.

2. Adjusting for number of publications

Figure 1 in the text documents the trend in specialization across successive cohorts of scientists using the AMS database, with the sample being restricted to scientists who have published more than one paper over the relevant span of time. As noted in the text, the cohort effects may be contaminated by concurrent trends in the number of papers published by the cohorts. A decline in the number of papers would mechanically lead to more specialization. Figure S1 illustrates the respective trend in the number of papers published. It is evident that there is no systematic trend in published output that would account for the cohort effects documented in Figure 1.

We also conducted a regression-based analysis as follows. We stacked the data of individual scientists across cohorts, and estimated a micro-level regression of the number of fields (or the Herfindahl index) on the number of papers published by the scientist over the respective time span. We then calculated the individual-level residuals from this regression, so that the residuals measure the adjusted index of specialization, netting out the impact of the number of papers. We then averaged this residual across cohorts. These averaged residuals are summarized in Figure S2, which again confirms the trend in cohort effects discussed in the text.

3. The impact of teamwork on specialization

The AMS database does not identify the coauthors of any given paper. But, as noted earlier, the merged AMS-ISI database can be used to identify all of the coauthors of each paper in the matched sample. We focus on the first 10 years of a mathematician's career in this matched sample, for mathematicians who first published between 1969 and 1999. For each mathematician, we construct a variable indicating if each paper he or she published in the first 10 years represents a contribution to a previously unexplored field. By construction, this variable must equal one for the first paper the mathematician ever publishes. We omit the first paper from the analysis that follows.

We stacked the data across mathematicians and papers, and estimate the regression model presented in Table S1, where the dependent variable is the indicator for whether the paper represents a contribution to a previously unexplored field (for that scientist). The two key independent variables indicate if the paper has two authors, or if the paper has at least three authors. The excluded variable indicates if the paper is solo-authored. Depending on the specification, the regression also includes a vector of fixed effects for the year of experience, a vector of person fixed effects, and a vector of year of publication fixed effects. The impact of teamwork (2 authors or more) is positive and statistically significant, regardless of the specification.

Table S3 considers three modified dependent variables. First, we construct an indicator for whether the paper was in a previously unexplored field *and* this particular author later contributed to this field again. Second, we construct an indicator for whether the paper was in a previously unexplored field *and* this particular author later contributed to this field again through his or her solo-authored work. Third, we construct an indicator for whether the paper was in a previously unexplored field *and* this particular author later contributed to this field again through his or her team-authored work. The key independent variables again measure the number of authors on the paper, where the excluded variable is the indicator for solo-authored papers. The regression coefficients indicate that the introduction of a new field through coauthored work leads to further related work by the author, including further solo-authored work.

4. Inter-country differences in specialization

The AMS database identifies the location of the mathematician systematically beginning only in 1984. For the cohort of mathematicians whose first paper of publication was between 1984 and 1990, we use the location of the first paper published in that period to allocate mathematicians across countries. We restrict the analysis to the 44 countries that had at least 50 mathematicians in the cohort. The analysis uses four variables: the average number of fields that mathematicians in each country published in over the first 20 years of their career; the average Herfindahl index across the mathematicians; the (log) number of mathematicians in each country who published at least one paper between 1984 and 1990; and the fraction of papers written by the cohort of mathematicians in each country that either have 2 authors or at least 3 authors (the excluded variable is the fraction of papers that are solo-authored.

Table S2 reports the regressions. All regressions are weighted by the size of the cohort in the country. The regressions show that the (log) number of mathematicians has a consistently negative effect on specialization, while the fraction of team-produced papers has a consistently positive effect on specialization across countries.

Fig. S1. Trends in papers published. The figure illustrates the trend in the average total number of papers published by each cohort in the respective time span.

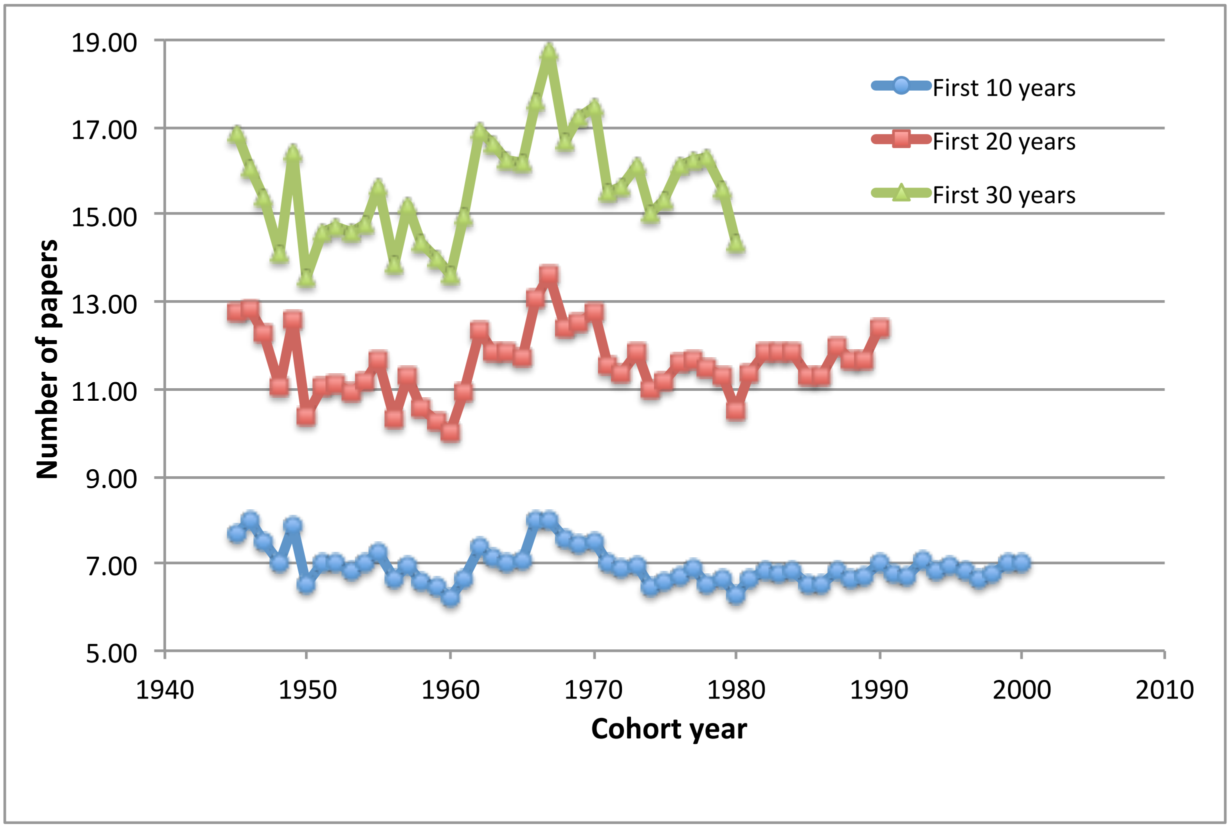


Fig. S2. The increase in specialization, holding constant number of papers across scientists. These figures present the average number of fields (A) or the average Herfindahl index (B) over the working life for scientists who first published in 1945-56, 1957-68, 1968-80, or 1981-92.

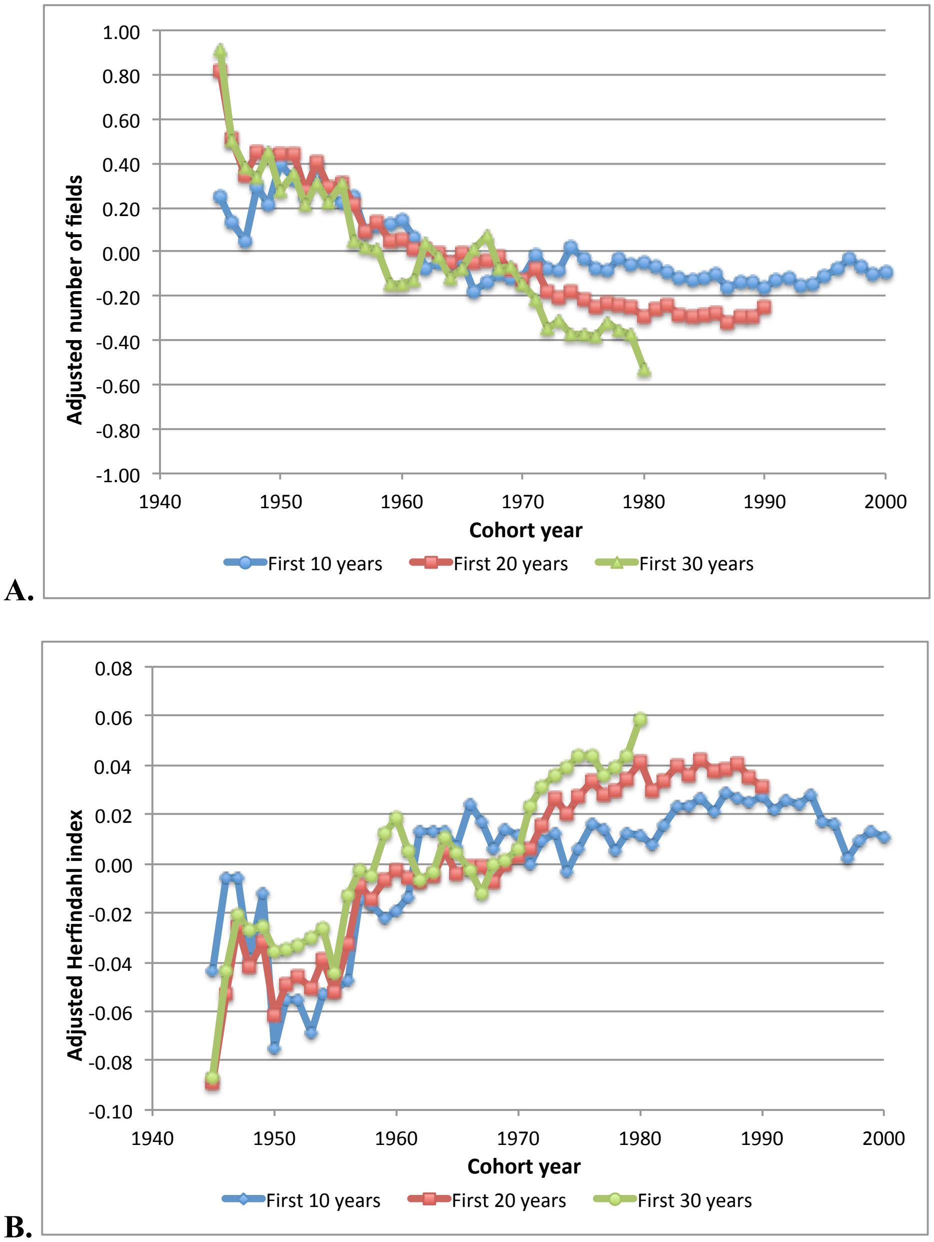


Table S1. The impact of coauthorships on the probability that a paper is in a new field

	Specification			
Independent variable:	(1)	(2)	(4)	(5)
Two authors	0.006	0.013	0.025	0.025
	(0.002)	(0.002)	(0.002)	(0.002)
Three or more authors	0.013	0.023	0.039	0.039
	(0.002)	(0.002)	(0.003)	(0.003)
Includes:				
Experience fixed effects	No	Yes	Yes	Yes
Person fixed effects	No	No	Yes	Yes
Year fixed effects	No	No	No	Yes

Notes: Robust standard errors are reported in parentheses and are clustered at the individual level. The experience fixed effects are a set of dummy variables indicating the year of experience (between 1 and 10) in which the paper was published. The year fixed effects are a set of dummy variables indicating the year of publication (between 1969 and 2008) in which the paper was published. The regressions have 404,980 observations.

Table S2. The relation between specialization, the size of the market, and the degree of specialization across countries

	Specification		
	(1)	(2)	(3)
Dependent variable = Number of fields			
Log number of mathematicians	-0.057	---	-0.036
	(0.020)		(0.018)
Fraction of papers with 2 authors	---	-1.153	-1.124
		(.133)	(1.838)
Fraction of papers with more than 2 authors	---	2.930	2.680
		(0.690)	(0.676)
R-squared	0.087	0.271	0.303
Dependent variable = Herfindahl index			
Log number of mathematicians	0.005	---	0.003
	(.002)		(.002)
Fraction of papers with 2 authors	---	0.215	0.213
		(.221)	(.230)
Fraction of papers with more than 2 authors	---	-0.246	-0.223
		(.071)	(0.073)
R-squared	0.052	0.189	0.210

Notes: Robust standard errors are reported in parentheses. The regressions are weighted by the size of the 1984-1990 cohort in the country. The regressions have 44 observations.

Table S3. The impact of coauthorships on the probability that a paper is in a new field *and* that this new field is later replicated by the author in either: (1) any of their work, (2) specifically in their solo-authored work, or (3) specifically in their team-authored work

	Dependent variable		
	(1) Paper in new field, and field is repeated later	(2) Paper in new field, and field is repeated later in solo-authored work	(3) Paper in new field and field is repeated later in team-authored work
Independent variable:			
Two authors	0.009	0.003	0.006
	(0.002)	(0.001)	(0.001)
Three or more authors	0.013	0.004	0.009
	(0.002)	(0.001)	(0.002)
Includes:			
Experience fixed effects	Yes	Yes	Yes
Person fixed effects	Yes	Yes	Yes
Year fixed effects	Yes	Yes	Yes

Notes: Robust standard errors are reported in parentheses and are clustered at the individual level. The experience fixed effects are a set of dummy variables indicating the year of experience (between 1 and 10) in which the paper was published. The year fixed effects are a set of dummy variables indicating the year of publication (between 1969 and 2008) in which the paper was published. The regressions have 404,980 observations.