Research output diversification and performance

Harris Dellas
Uni Surrey, St. Louis FED, Karl Brunner Institute

Manu García, Carlos Garriga, Christian Zimmermann
Federal Reserve Bank of St. Louis

July 2025

Research performance

$$Y = F(A, X(A, Z, ..), L)$$

Research performance, **Y**, depends on a researcher's

- *innate* characteristics (ability,..): **A**
- *choices* (effort, co-authors, goodness of fit with chosen field,..), **X**. Some of these choices may vary with –typically unobserved- A, which creates an identification problem
- "*luck*" (editors' tastes/"proximity," fads,..): **L**

We study the role of the **choice** of the degree of research **output** variety (an element of X)

We estimate its effect as well as its explanatory power (conditional on A)

Measures of research performance

- Publications and citations (in various forms) are the most commonly used measures
- A debate about which indicator better captures productivity; and for each indicator, which of its alternative representations is best
  (A "thin" literature: Hamermesh, 2018, Perry-Reny, 2016, Heckman-Moktan, 2020,...)
- We ignore this debate, use both measures.

Determinants of research performance

- The literature on the determinants of research performance *in economics* is quite thin
- CONNECTIONS: Brogaard-Engelberg-Parsons, 2014, Laband–Piette, 1994
- PEDIGREE: Conley-Onder, 2014
- Similarly thin *outside economics*
  Dubois-Rocher-Schenkler, 2013 (very comprehensive)

Determinants of research performance

- An even "thinner" literature on the determinant this paper focuses on (degree of research output specialization)
- Pretty much nothing in economics
- A bit of relevant work
  in the sciences (Abramo- D'Angelo-Di Costa, 2017, 2019, Kelchtermans- Neicu-Veugelers, 2022)
  and mathematics
  (Dubois-Rocher-Schenkler, 2013)

Questions

- Does the choice of the degree of specialization in research output matter for performance (level/quality of output)?
- Does its time profile (early vs late in the career) make a difference?
- If output variety is rewarded, are there any simple ways to pursue it?

Specialization-productivity: A bit of theory

- Trade theory. With complete asset markets, perfect output specialization maximizes the value of production when the production function is IRS or CRS.
- With incomplete markets, insurance motives dictate imperfect specialization and a sacrifice of output.
- But research production may have features that push in the opposite direction (e.g. TFP improvement from either cross-fertilization or joint work with people from other fields).

An example

In the Manhattan project, Oppenheimer instituted a weekly colloquium for hundreds of scientists. He understood the value of gathering people from different parts of a project in the same place, encouraging them to discuss their work and combine their ideas.

As Bethe commented, ''*very often a problem discussed in one of these meetings would intrigue a scientist in a completely different branch of the laboratory, and he would come up with unexpected solutions.*"

The RePEc data

The key variable: The degree of output specialization

Constructed from the RePEc data set

The RePEc data

- Who authored, which paper, (working paper/journal article/book chapter), where it was published, who reads it and where it is cited
- It contains info on various attributes of researchers (pedigree, academic age, gender, affiliation,..)
- It has several categories/subcategories of performance (which are combined to produce a ranking of authors)

Performance

- Performance depends on both quantity and quality
- The quality of a publication/citation is determined from the *impact factor* of the publishing/citing venue (journal)
- Impact factors are critical and controversial
- The RePEc impact factors are determined internally within the RePEc data set and evolve in real time

Indicators of author performance

- Quality adjusted pages: the number of journal pages, weighted by the *recursive* impact factor of the publishing journal
- # of top 5 publications.
- Quality adjusted citation count: the number of citations, excluding self-citations, weighted by the recursive impact factor of the citing venue
- The number of distinct works, each weighted by the recursive impact factor
- RePEc author rank (based on publications, citations and many other criteria)
- number of NEPCites

Recursive impact factor: similar to the Google Page-Rank which ranks web pages higher if they are linked to many others, even more so if the web sites they are linked to have a high Page-Rank themselves; RePRc ranks each journal and working paper series rather than each web page

Measuring research output diversification

- The **JEL classification system**: 20 primary categories (a letter), each with secondary (adds 1 digit), for a total of 141, and tertiary subcategories (adds a 2nd digit), for a total of 867
  E.g **F** is International Economics); **F1** is International Trade. **F18** is Trade and Environment
- A diversification index for each classification category
- Our baseline is JEL tertiary (all inclusive)
- A portfolio of works is simply a portfolio of JEL codes
- For robustness purposes we also use a diversification index based on the **NEP classification system**. It has 100 fields and covers working papers only.

Specialization Diversification Index (SDI)

Let i be the number of papers that contain code j. SDI is constructed as

$$SDI = \frac{\sum_j i^2}{(\sum_j i)^2}$$

A lower value of SDI represents greater output specialization. The larger the number of JEL fields in one's portfolio of works and the more even their distribution, the lower one's SDI score and degree of specialization.

Similar to the Herfindahl−Hirschman Index (HHI) with the researcher's portfolio replacing the market, and the individual JEL codes the firms: instead of the distribution of firm market shares, we have the shares of JEL codes in the portfolio.

Outline: Establishing stylized facts

- The relation between diversification and performance in the cross section (author rank, research output –pubs, cites)
- Joint work, diversification, research output; quality vs quantity
- The time profile of diversification and research output: early diversification and short/ long term performance/ research attrition

▸ stats

Outline: Establishing causality

Diversification and research output: Filtering out the effect of unobserved ability

- Subset of data, set of authors with top 5 pubs
- partitions on the basis of pedigree, performance percentiles
- fixed effects in panel data
- IV estimation

RePEc rank and diversification

Two main findings:

- There is a positive, statistically and economically significant association between rank and diversification
- As the measure of diversification becomes broader, its association with rank diminishes: The estimated elasticity and its explanatory power decline
- Similar results obtain for gender and age specific subsets of the data

▸ sdirank

Diversification and productivity

RePEc rank can be well (80%) accounted for by a couple of criteria: quality adjusted pages, quality adjusted citations, and # of NEP citing fields (breadth)

▸ rank

Each of these determinants bears a significant relationship to diversification

▸ perf

Joint work and performance

Joint work increases

- the amount of labor input in the research production function (more people working on the same paper)
- production TFP due to complementarities in the skills and expertise of co-authors, allowing greater input specialization within the team according to CA

Two variables

- Stock of distinct co-authors
- Average # of co-authors per paper
- Findings:
  1. Joint work matters

joint

  2. It matters more for output quality than for output quantity

joint1

joint2

Early diversification and performance

SDI: (a) Baseline; (b) The number of distinct codes JEL codes in one's portfolio; (c) the share of distinct JEL codes during the first 5 years in lifetime number of distinct codes (15 years)
Performance: Research output and research attrition (the cessation of publication)

- Those who have written enough papers of sufficient variety during this early window have fared better, in particular in the longer term
- This is true for all three measures of diversification

early1
early2
attr1
attr2

Causality: Attempting to control for ability

- Conditioning on top 5 pubs

top5a

- Conditioning of percentile of performance distribution, pedigree, etc.

percent

Statistical and economic significance remains strong (the latter a bit diminished because of the –partial – accounting of the ability-diversification interaction)

Causality: Attempting to control for ability, PANEL data

- Full data set, panel
- Fixed effects –FE– should take care of the –constant– cross sectional variation in ability
- The FE estimation suggests a statistically and economically (within $R^2 = 0.14$) significant effect of diversification on research performance
- Effect smaller than under OLS
- But there may exist (?) time varying cross section variation in ability that matters for the choice of specialization
- We use IV estimation with the average SDI of one's co-authors as an instrument

panel1

- Set of authors with top 5 publication(s), panel Encouragingly, similar, even stronger results

panel2

Conclusions

- An empirical measure of the degree of specialization of a portfolio of works
- Strongly related to standard measures of performance
- Joint work leads to greater output variety/quality
- The time profile of diversification matters for performance: early diversification is associated with better short and long term performance and a lower rate of research attrition.

Conclusions

- The fixed effects-IV estimation as well as the other exercises that attempt to control for unobserved ability support the claim that the associations reported above may represent a causal relationship running from the individual's choice of output variety to her performance.
- This choice represents a consequential (statistically and economically significant) part of one's research strategy.

TABLE: RePEc Data: Descriptive Statistics

Variable	Mean	SD	Min	P10	Median	P90	Max
Items	24.94	38.27	1.00	2.00	12.00	61.00	1,427
# Articles	14.58	24.11	0.00	0.00	7.00	36.00	762
WScPages	109.1	335.2	0.00	0.00	9.21	269.5	12,830
WScCites	152.3	744.2	0.00	0.00	7.83	250.9	36,307
JEL	19.13	21.12	0.00	3.00	12	44	332
*JEL_L*1	11.61	10.45	0.00	2.00	9	25	108
JEL_L	5.59	3.63	0.00	2.00	5	11	20
Reg. Coauth	9.32	12.92	1.00	1.00	5.00	21	291
Coauthors pp	3.29	1.72	0.07	1.56	3.00	5.3	46
SDI	0.17	0.18	0.01	0.04	0.11	0.33	1.00
*SDI_L*1	0.40	0.23	0.08	0.18	0.33	0.72	1.00
SDI_L	0.25	0.21	0.02	0.08	0.18	0.50	1.00
SDI_NEP	0.27	0.23	0.00	0.09	0.18	0.50	1.00
SDI_coauth	0.09	0.07	0.01	0.04	0.07	0.15	1.00

The relation between diversification and performance

Table: Diversification and RePEc Author Score (Rank)

	Log Score s				Log Score			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
(c)	−11.547***	−11.027***	−10.498***	−11.272***	−11.923***	−11.570**hlien*	−11.238***	−11.909
	(0.010)	(0.010)	(0.010)	(0.013)	(0.009)	(0.009)	(0.009)	(0.012
SDI	−0.705***				−0.545***			
	(0.004)				(0.004)			
SDI_L		−0.622***				−0.458***		
		(0.006)				(0.005)		
SDI_L1			−0.482***				−0.346***	
			(0.008)				(0.007)	
SDI_NEP				−0.846***				−0.692*
				(0.007)				(0.006
Age					0.041***	0.046***	0.050***	0.049**
					(0.000)	(0.000)	(0.000)	(0.000
R^2	0.307	0.167	0.059	0.231	0.501	0.419	0.361	0.495
# obs.	58,003	58,003	58,003	44,794	58,002	58,002	58,002	44,794

table1

A few measures of research output explain score (rank) well

Table: Research Output and Score (Rank)

	Score
$LWScPages$	0.0693^{***}
	(0.0016)
$LWScCites$	0.0586^{***}
	(0.0022)
$NEPCites$	0.0230^{***}
	(0.0002)
R^2	0.8100
Num. obs.	45995

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$.

table2

TABLE: Diversification and Research Output: QAP, QAC, NEPcites

	LWScPages		*LWScCites*		*NEPCites*	
	(1)	(2)	(3)	(4)	(5)	(6)
(c)	−0.122***	−1.131***	−0.278***	−1.501***	−9.436***	−22.912***
	(0.034)	(0.034)	(0.035)	(0.033)	(0.375)	(0.354)
SDI	−1.308***	−1.115***	−1.327***	−1.093***	−18.631***	−16.054***
	(0.014)	(0.013)	(0.014)	(0.013)	(0.151)	(0.136)
Age		0.075***		0.091***		1.001***
		(0.001)		(0.001)		(0.009)
R^2	0.164	0.283	0.164	0.335	0.249	0.409
# obs.	45995	45995	45995	45995	45995	45995

table2

Joint work and performance

Table: Joint work and research output (Cohorts 2000–2005)

	LWScCites		LWScPages	
	(1)	(2)	(3)	(4)
(c)	2.593***	−0.064	2.604***	0.718**
	(0.029)	(0.253)	(0.029)	(0.258)
Coauthors	0.063***	0.055***	0.059***	0.046***
	(0.002)	(0.002)	(0.002)	(0.002)
Age		0.128***		0.088***
		(0.013)		(0.013)
Items		0.005***		0.008***
		(0.001)		(0.001)
R^2	0.187	0.203	0.160	0.178
Num. obs.	7,691	7,691	7,691	7,691

TABLE: Joint work and performance (Quality un-adj. pages)

	(1)	(2)	(3)	(4)	(5)
SDI	−0.903***		−0.414***		−0.890***
	(0.006)		(0.007)		(0.007)
Stock Reg. Coauth		0.853***	0.651***		
		(0.005)	(0.006)		
Avg. Coauth pp				0.474***	0.095***
				(0.013)	(0.011)
R^2	0.284	0.393	0.432	0.025	0.285
# obs.	52,487	52,918	52,487	52,918	52,487

table3

TABLE: Co-Authorship and performance (Quality adj.pages)

	(1)	(2)	(3)	(4)	(5)
(c)	-1.305^{***}	-0.663^{***}	-1.397^{***}	0.372^{***}	-2.249^{***}
	(0.031)	(0.019)	(0.027)	(0.030)	(0.036)
SDI	-1.653^{***}		-0.532^{***}		-1.482^{***}
	(0.013)		(0.014)		(0.013)
Stock R. Coauth		1.749^{***}	1.492^{***}		
		(0.009)	(0.011)		
Avg coauth				1.833^{***}	1.213^{***}
				(0.025)	(0.023)
R^2	0.231	0.403	0.419	0.093	0.269
# obs.	52,487	52,918	52,487	52,918	52,487

Table: Early diversification and short- and long-run performance, Quality adj. pages–citations (Cohorts 2000–2005)

	Pages after 5y		Pages after 15y		Cites after 15y	
SDI 5ys	−0.484***	−0.125**	−0.736***	−0.432***	−0.976***	−0.567***
	(0.040)	(0.042)	(0.037)	(0.039)	(0.042)	(0.043)
Items 5ys		0.046***		0.039***		0.052***
		(0.002)		(0.002)		(0.002)
R^2	0.019	0.079	0.049	0.098	0.067	0.135
# obs.	7,691	7,691	7,691	7,691	7,691	7,691

▸ table4

Table: Early diversification and short/long run performance, cohorts 2000-2005: Alternative diversification measure (distinct # JELs)

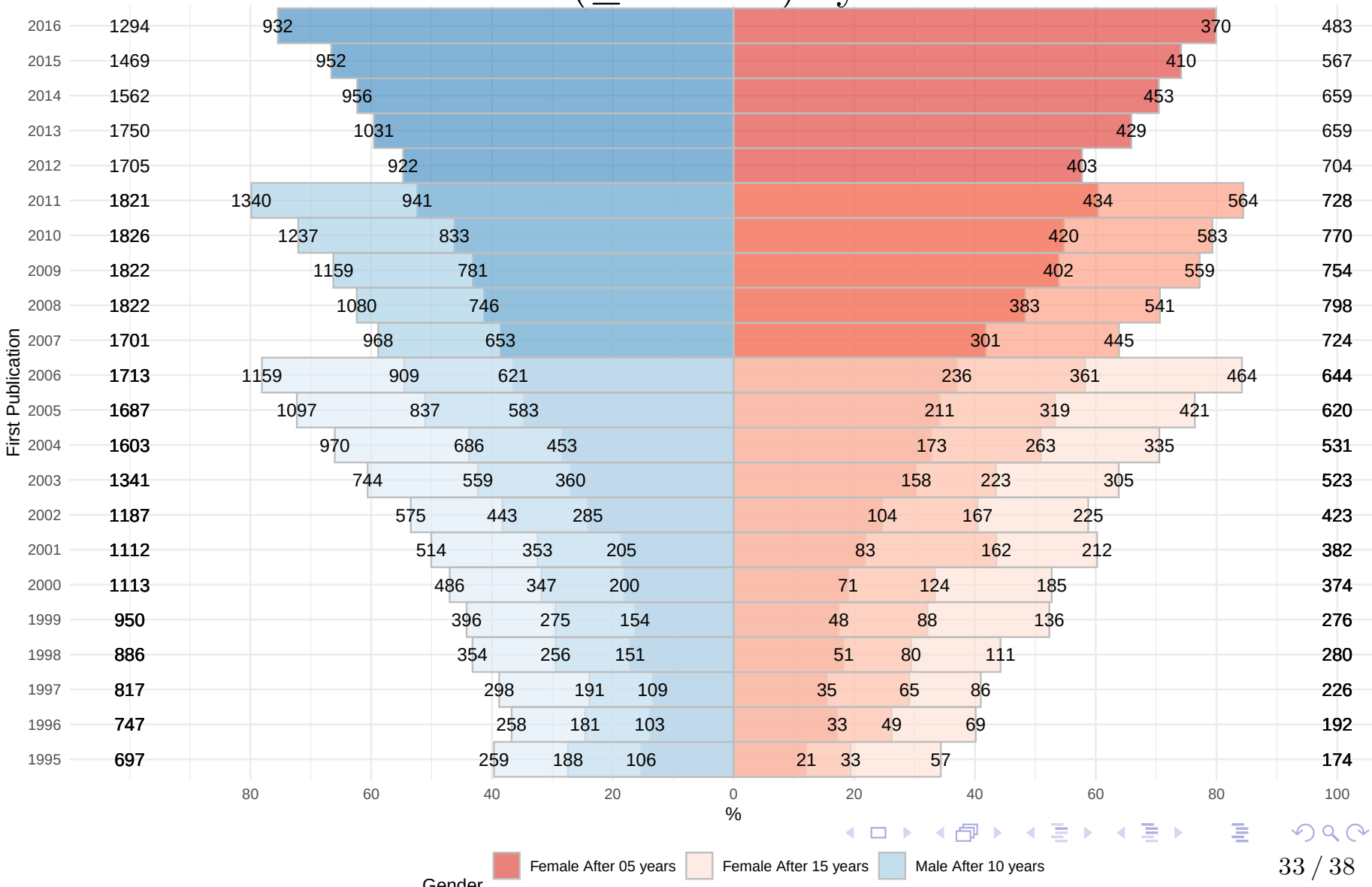
	Pages 5ys	Pages 5ys	Pages 15ys	Pages 15ys	Cites 15ys	Cites 15ys
#JEL 5ys	0.043***	0.003	0.060***	0.028***	0.082***	0.041***
	(0.003)	(0.004)	(0.003)	(0.003)	(0.003)	(0.004)
Items 5ys		0.047***		0.037***		0.048***
		(0.002)		(0.002)		(0.002)
R^2	0.026	0.078	0.057	0.093	0.083	0.130
# obs.	7691	7691	7691	7691	7691	7691

table4

Early diversification and research attrition

table4

Figure: Research Attrition (≤ 2 items) by Cohort and Gender



$$P(out_t = 1) = F(\beta_0 + \beta_1 SDI\_5 + \beta_2 items_5 + \varepsilon_i)$$

Table: Attrition (Zero Output) after 5, 10, 15 Years (Cohorts 2000–05)

	5-Year Attrition		10-Year Attrition		15-Year Attrition	
	(1)	(2)	(3)	(4)	(5)	(6)
(c)	−0.91***	−0.50***	−0.05	−0.01	0.72***	0.45***
	(0.13)	(0.07)	(0.09)	(0.05)	(0.08)	(0.05)
SDI 5ys	0.92***	0.21***	0.71***	0.21***	0.59***	0.24***
	(0.07)	(0.04)	(0.05)	(0.03)	(0.04)	(0.03)
Items 5ys		−0.06***		−0.03***		−0.02***
		(0.01)		(0.00)		(0.00)
AIC	3877.95	3650.85	7589.15	7330.27	10141.34	9994.25
BIC	3891.85	3671.70	7603.05	7351.11	10155.24	10015.09
Log Lik	−1936.98	−1822.43	−3792.58	−3662.13	−5068.67	−4994.13
Num. obs.	7,691	7,691	7,691	7,691	7,691	7,691

table4

Another performance indicator: Top 5 pubs

Table: Distribution of Authors, Top-5 Journal Publications

Articles	Authors	Articles	Authors	Articles	Authors	Articles	Authors
1	2,871	11	58	21	14	31	7
2	1,186	12	66	22	11	32	4
3	631	13	43	23	11	33	3
4	353	14	38	24	10	34	5
5	303	15	34	25	9	35	2
6	234	16	31	26	4	36	1
7	164	17	28	27	7	37	2
8	114	18	25	28	3	38	6
9	96	19	20	29	3	39	2
10	79	20	14	30	5	40+	16

top5b

Diversification and Output: Top 5 pubs

Table: Diversification and Top-5 Pubs (Only top 5 authors)

	Number of Top-5 Publications		
	(1)	(2)	(3)
SDI	-1.777^{***}	-2.025^{***}	-1.214^{***}
	(0.095)	(0.089)	(0.092)
Academic Age		0.137^{***}	0.069^{***}
		(0.004)	(0.005)
Number of Articles			0.049^{***}
			(0.002)
R^2	0.050	0.176	0.241
Num. obs.	6,617	6,617	6,617

table5

TABLE: Lscore, SDI, number of items- Top 10 school graduates, 70-100th percentile of pages

Period	n	R-sq	SDI	Items
1981-1983	34	0.41	0.28	0.01***
1984-1986	40	0.45	-0.53***	0.01***
1987-1989	56	0.50	-0.27	0.01***
1990-1992	51	0.27	-0.39**	0.01***
1993-1995	57	0.46	-0.78***	0.01***
1996-1998	62	0.52	-0.61***	0.01***
1999-2001	73	0.39	-0.44***	0.01***
2002-2004	71	0.52	-0.50***	0.02***
2005-2007	62	0.39	-0.40**	0.02***
2008-2010	59	0.20	-0.39**	0.01**
2011-2013	39	0.44	-0.32**	0.03***
2014-2016	31	0.69	-0.43***	0.03***

Diversification and performance: Panel

Table: SDI and performance: Fixed effects, IV (Full Sample)

	OLS	OLS + TE	OLS + FE	IV	IV + TE	IV + FE
(c)	-0.168***			2.620***		
	(0.010)			(0.032)		
SDI	-1.501***	-1.785***	-0.787***			
	(0.004)	(0.023)	(0.009)			
SDI				-0.305***	-2.503***	-1.195***
				(0.015)	(0.061)	(0.047)
Num.Obs.	776379	776379	776379	725973	725973	725973
R2	0.137	0.202	0.962	0.049	0.140	0.956
R2 Within		0.184	0.132		0.123	0.099
RMSE	2.82	2.72	0.59	2.81	2.67	0.60
Std.Errors	IID	by: Year	by: ID	IID	by: Year	by: ID
FE: Year		X	X		X	X
FE: ID			X			X

Diversification and performance (top 5): Panel

Table: SDI and Performance: Fixed effects and IV (Top 5)

	OLS	TE	FE + TE	IV	TE	FE + TE
(c)	3.893***			4.394***		
	(0.013)			(0.026)		
SDI	-1.000***	-1.148***	-0.687***			
	(0.005)	(0.015)	(0.017)			
SDI				-0.799***	-1.630***	-0.951***
				(0.011)	(0.065)	(0.075)
Num.Obs.	145617	145617	145617	143745	143745	143745
R2	0.198	0.216	0.904	0.183	0.165	0.899
R2 Within		0.203	0.144		0.154	0.121
RMSE	1.57	1.55	0.54	1.55	1.57	0.55
Std.Errors	IID	by: Year	by: ID	IID	by: Year	by: ID
FE: Year		X	X		X	X
FE: ID			X			X

table7