Online Appendix

Who Becomes an Inventor in America? The Importance of Exposure to Innovation

Alex Bell, Harvard University
Raj Chetty, Harvard University and NBER
Xavier Jaravel, London School of Economics
Neviana Petkova, Office of Tax Analysis, US Treasury
John Van Reenen, MIT and Centre for Economic Performance

November 2018

Matching Algorithm

The patent data were linked to the tax records using a variant of the matching algorithm developed in Chetty et al. (2014a) to link the New York City school district data to the tax records. Chetty et al. (2011) show that the match algorithm outlined below yields accurate matches for approximately 99% of cases in a school district sample that can be exactly matched on social security number. Note that identifiers were used solely for the matching procedure. After the match was completed, the data were de-identified (i.e., individual identifiers such as names were stripped) and the statistical analysis was conducted using the de-identified dataset.

Before beginning the match process, the names were standardized as follows. First, suffixes sometimes appear at the end of taxpayers' first, middle, or last name fields. If these fields end with a space followed by "JR", "SR", or a numeral I-IV, the suffix is stripped out and stored separately from the name. Second, the USPTO database separates inventor names into "first" and "last," but the tax data often separates names into first, middle, and last. In practice, many inventors include a middle initial or name in the first name field. Whenever there is a single space in the inventor's first name field, for the purposes of matching, we allow the first string to be an imputed first name, and the second string to be an imputed middle name or initial. The use of these imputed names is described below.

The matching algorithm proceeds in seven steps. Inventors enter a match round only if they have not already been matched to a taxpayer in an earlier round. Each round consists of a name criterion and a location criterion. The share of data matched in each round is documented below.

- **Stage 1**: Exact match on name and location.
  - Name match: The inventor's last name exactly matches the taxpayer's last name. Either the inventor's first name field exactly matches the concatenation of the IRS first and middle name fields or the IRS middle name field is missing, but the first name fields match. If an imputed middle name is available for the inventor, candidate matches are removed if they have ever filed at the IRS with a middle name or initial that conflicts with the inventor's.
  - Location match: The inventor's city and state must match some city and state reported by that taxpayer exactly.
  - 49% of patents are uniquely matched in this stage.

- **Stage 2**: Exact match on imputed name data and location.
  - Name match: The inventor's last name exactly matches the taxpayer's last name and the taxpayer's last name is the same as the inventor's imputed first name. Either the inventor's imputed middle name/initial matches one of the taxpayer's middle/initial name fields, or one of the two is missing. For inventors with non-missing imputed middle names, priority is given to matches to correct taxpayer middle names rather than to taxpayers with missing middle names. As above, candidate matches are removed if they have ever filed at the IRS with a conflicting middle name or initial.
  - Location match: As above, the inventor's city and state must exactly match some city and state reported by that taxpayer.
  - 12% of patents are uniquely matched in this stage.

- **Stage 3**: Exact match on actual or imputed name data and 1040 zip crosswalked.
  - Name match: The inventor's last name exactly matches the taxpayer's last name. The inventor's first name matches the taxpayer's first name in one of the following situations, in order of priority: (1) inventor's first name is the same as the taxpayer's combined first and middle name; (2) inventor's imputed first name matches taxpayer's and middle names match on initials; (3) inventor has no middle name data, but inventor's first name is the same as the taxpayer's middle name.
  - Taxpayers are removed if they are ever observed filing with middle names in conflict with the inventor's.
  - Location match: The inventor's city and state match one of the city/state fields associated with one of the taxpayer's 1040 zip codes.
  - 3% of patents are uniquely matched in this stage.

- **Stage 4**: Same as previous stage, but using names from 1040 forms instead of names from W-2 forms.
  - Name match: The inventor's name matches the name of a 1040 (or matches without inventor's middle initial/name and no taxpayer middle initials/names conflict with inventor's).
  - Location match: The inventor's city and state must match some city and state reported by that taxpayer exactly.
  - 6% of patents are uniquely matched in this stage.

- **Stage 5**: Match using W-2 full name field.
  - Name match: The inventor's FULL name exactly matches the FULL name of a taxpayer on a W2.
  - Location match: The inventor's city and state match one of the city/state fields associated with one of the taxpayer's 1040 zip codes.
  - 8% of patents are uniquely matched in this stage.

- **Stage 6**: Fuzzy match using W-2 full name field.
  - Name match: The inventor's full name (minus the imputed middle name) exactly matches the full name of a taxpayer on a W2.
  - Location match: The inventor's city and state match one of the city/state fields associated with one of the taxpayer's 1040 zip codes.
  - 1% of patents are uniquely matched in this stage.

- **Stage 7**: Match to all information returns.
  - Name match: The inventor's full name exactly matches the full name of a taxpayer on any type of information return form.
  - Location match: The inventor's city and state match one of the city/state fields associated with one of the taxpayer's information return forms.
  - 6% of patents are uniquely matched in this stage.

References

R. Chetty, J. N. Friedman, N. Hilger, E. Saez, D. W. Schanzenbach, and D. Yagan. How Does your Kindergarten Classroom Affect your Earnings? Evidence from Project STAR. *The Quarterly Journal of Economics*, 126(4):1593–1660, 2011.

R. Chetty, J. N. Friedman, and J. E. Rockoff. Measuring the Impacts of Teachers I: Evaluating Bias in Teacher Value-Added Estimates. *American Economic Review*, 104(9):2593–2632, 2014a.

APPENDIX TABLE I

Association Between Patent Rates and Upper-Tail Incomes with 3rd grade Math vs. English Test Scores

Dependent Variable:	Inventor (per 1,000 Individuals)					In Top 1% of Income Distribn. (per 1,000 Individuals)		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
3rd Grade Math Score (SD)	0.85***		0.85***	0.91***		11.07***		8.08***
	(0.09)		(0.13)	(0.13)		(0.36)		(0.46)
3rd Grade English Score (SD)		0.68***	0.05		0.08		10.19***	4.15***
		(0.09)	(0.11)		(0.12)		(0.34)	(0.42)
Fixed Effects	None	None	None	English Ventile	Math Ventile	None	None	None
Mean of Dependent Variable	0.83	0.86	0.86	0.86	0.86	9.68	9.72	9.84
Observations	222,369	214,265	209,722	209,722	209,722	165,422	161,275	158,016

Notes: This table examines the extent to which third grade test scores are predictive of patent rates and upper-tail earnings outcomes. The sample in columns 1-5 consists of children in the 1979-1985 birth cohorts who attended New York City public schools in 3rd grade and were linked to the tax data. The sample in columns 6-8 consists of children who appear in both the NYC school district and intergenerational samples (1980-84 birth cohorts). Each column shows the coefficients and robust standard errors (in parentheses) from a separate OLS regression run at the student level; *** denotes $p < 0.001$. In Columns 1-5, the dependent variable is an indicator for being an inventor, defined as applying for a patent between 2001-2012 or being granted a patent between 1996-2014. In columns 6-8, it is an indicator for being in the top 1% of the individual income distribution in 2012 when compared to other individuals in the same birth cohort in the NYC school district sample. The dependent variables in each column are math and English test scores in 3rd grade. Test scores, which are based on standardized tests administered at the district level, are normalized to have mean zero and standard deviation one by year and grade. In Columns 4 and 5, we control for English and math scores non-parametrically using ventile fixed effects (20 bins) rather than a linear control. In all columns, coefficients are scaled so that they can be interpreted as the effect of a 1 SD change in test scores on the number of individuals per 1,000 who have the relevant outcome.

APPENDIX TABLE II

Fraction of Gender Gap in Innovation Explained by Differences in Test Scores

A. Percent of Innovation Gap Explained by 3rd Grade Math Test Scores

	Patent Rates for Women	Patent Rates for Men	Gender Gap
	(1)	(2)	(3)
Raw Estimates	0.43	1.13	0.70
	(0.06)	(0.09)	(0.11)
Reweighted to Match 3rd Grade Scores of Men	0.45	1.13	0.68
	(0.06)	(0.09)	(0.11)
	Gap in Innovation Explained by 3rd Grade Test Scores:		2.4%

B. Percent of Gap Explained by Test Scores Grades 3-8

Grade	Percent of Innovation Gap Explained by Math Test Scores in Grade *g*
3	2.4%
	(0.4)
4	2.4%
	(0.3)
5	3.4%
	(0.4)
6	4.6%
	(0.5)
7	6.8%
	(0.8)
8	8.5%
	(1.0)
Slope:	1.3
	(0.2)

Notes: This table shows how much of the gender gap in patent rates can be explained by test scores using the New York City school district sample. Panel A is constructed in exactly the same way as Table II, comparing girls with boys instead of low-income children with high-income children. Panel B presents estimates of the gender gap in innovation that can be explained by test scores in grades 3-8, analogous to the estimates in Figure V. The slope estimate reported at the bottom is estimated using an OLS regression of the six estimates on grade.

APPENDIX TABLE III

Distance Between Technology Classes: Illustrative Example

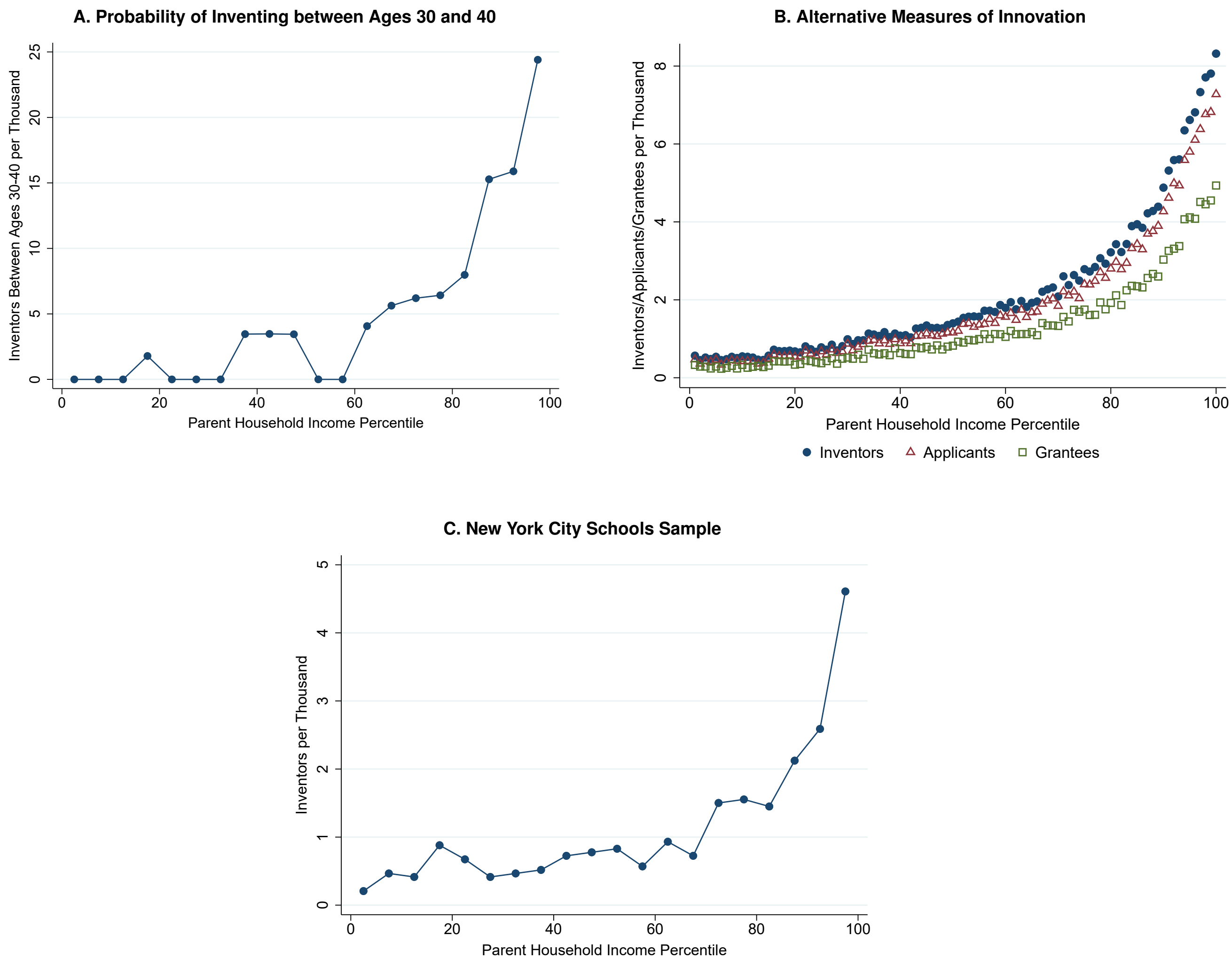
Category: Computers + Communications

Sub-category: *Communications*

Technology Class (= 375)	Distance Rank (*d*)
Pulse or Digital Communications	0
Demodulators	1
Modulators	2
Coded Data Generation or Conversion	3
Electrical Computers: Arithmetic Processing and Calculating	4
Oscillators	5
Multiplex Communications	6
Telecommunications	7
Amplifiers	8
Motion Video Signal Processing for Recording or Reproducing	9
Directive Radio Wave Systems and Devices (e.g., Radar, Radio Navigation)	10

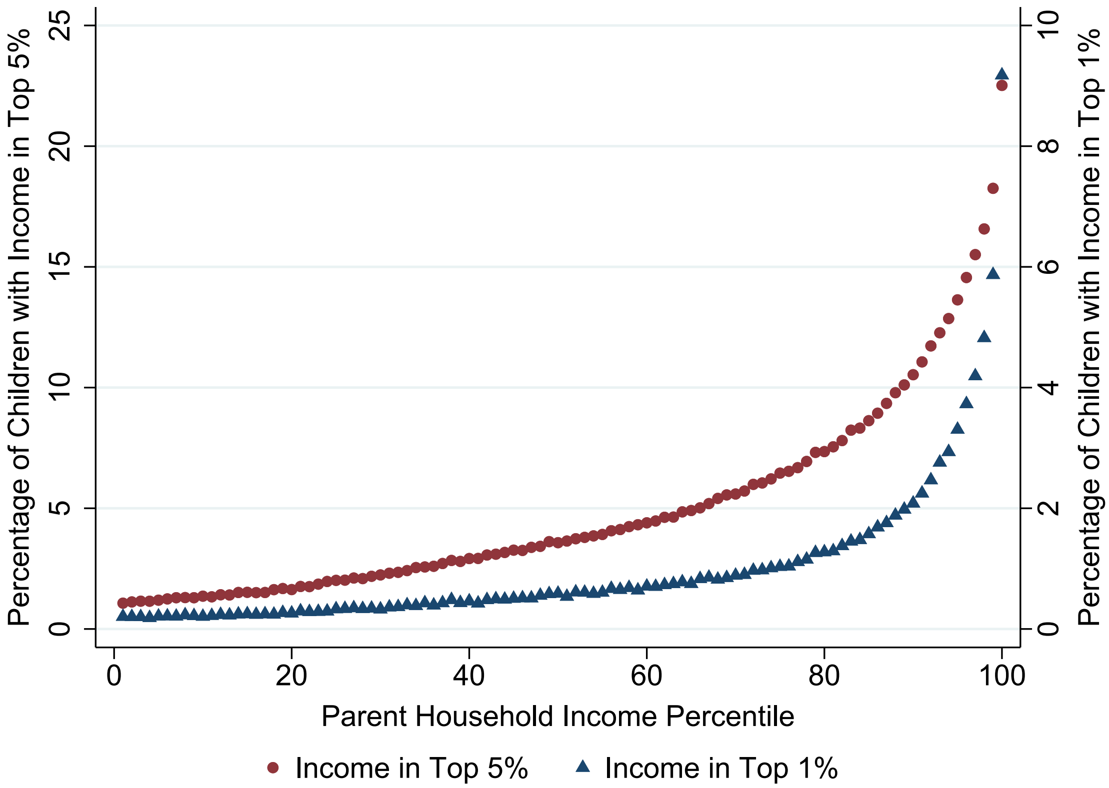
Notes: This table provides an example of our measures of distance between technology classes. We define the distance between two technology classes A and B by computing the share of inventors in class A who also invent in class B; the higher the share of common inventors, the lower the distance between A and B. We convert this distance metric to an ordinal measure, defining d=0 for the own class, d=1 for the next nearest class, etc. The table lists the 10 closest classes to the "Pulse or Digital Communtications" class, which falls within the Communications subcategory of the Computers + Communications category.

ONLINE APPENDIX FIGURE I: Patent Rates vs. Parent Income: Sensitivity Analysis



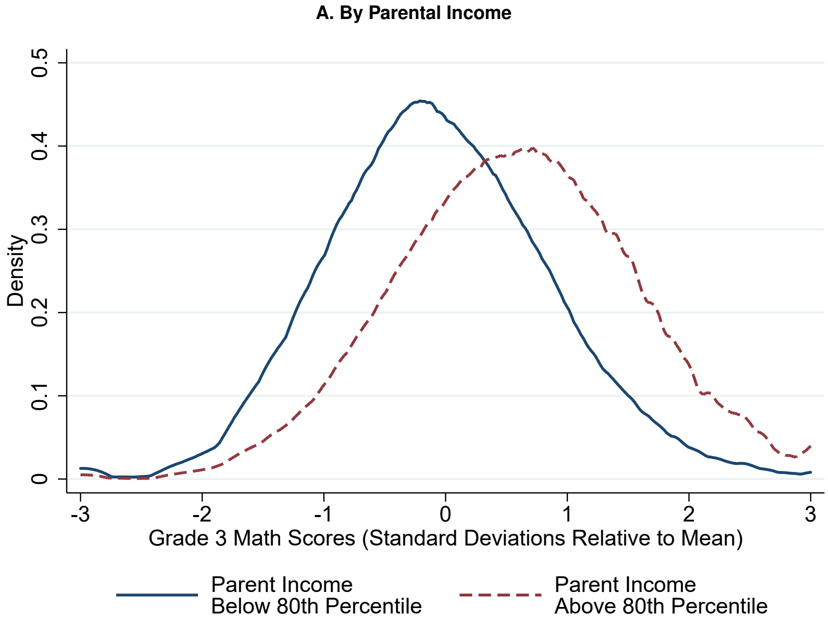
Notes: This figure replicates Figure Ia using alternative samples and definitions. Panel A uses data from the 1971-72 birth cohorts in the Statistics of Income sample, a 0.1% sample of tax returns (see Section II.B for details). This sample allows us to examine whether an individual filed a patent application or was granted a patent between the ages of 30 and 40, rather than just by their early thirties as in Figure Ia. Given the smaller sample size, we plot mean patent rates between the ages of 30-40 by parent ventile (20 bins) rather than percentiles in this figure. In Panel B, the series in circles replicates Figure Ia exactly, where inventors are defined as those who applied for a patent between 2001-2012 or were granted a patent between 1996-2014. The other two series in that figure show the fraction of individuals who applied for patents and the fraction who were granted patents separately. Panel C replicates the baseline series in Figure Ia (plotting the fraction of inventors) using the subset of children in the New York City public schools sample. In this figure, we rank parents within the NYC sample based on their household incomes and plot the fraction of children who become inventors by 2014 by parent income ventiles.

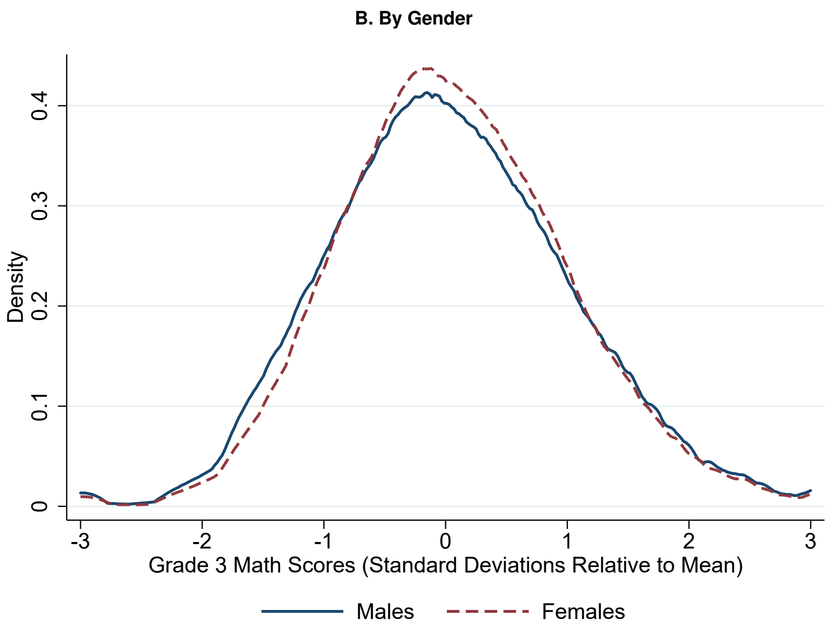
ONLINE APPENDIX FIGURE II:
Fraction of Children with Incomes in Upper Tail vs. Parent Income



Notes: This figure replicates Figure I, replacing the outcome variable with an indicator for having mean individual income in 2011-12 in either the top 1% or top 5% of the income distribution among individuals in the same birth cohort. The sample is our core intergenerational sample of the 1980-84 birth cohorts.

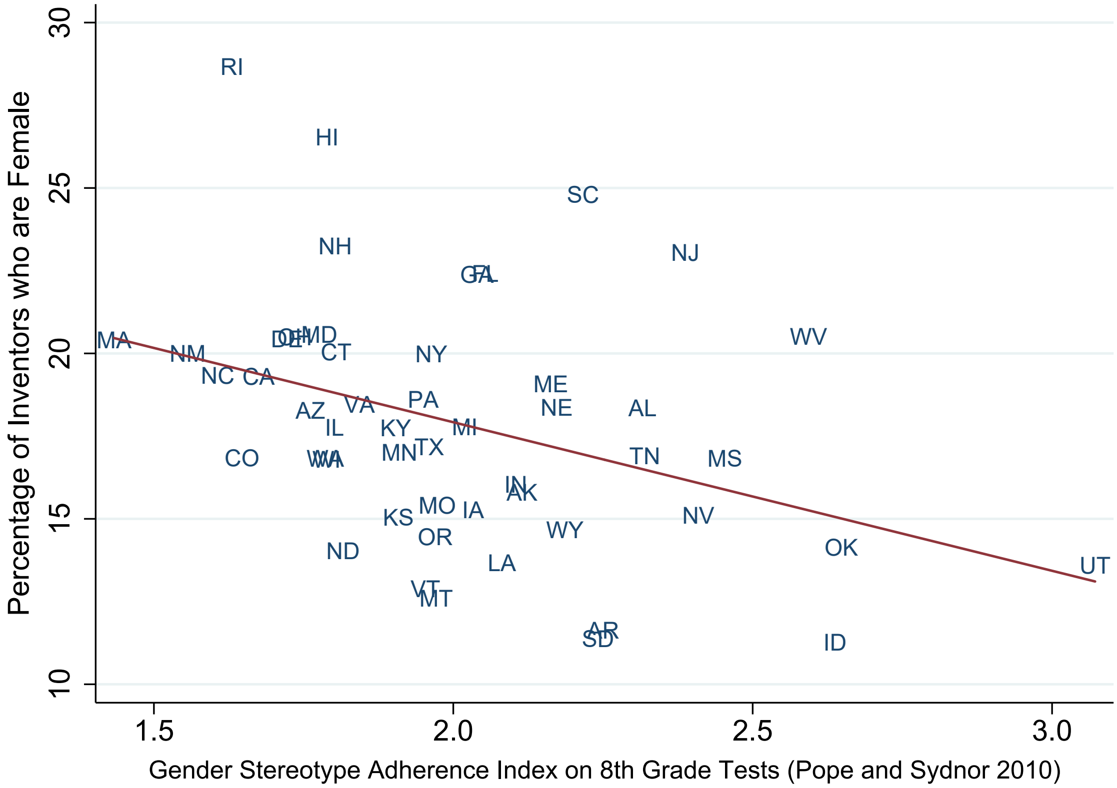
ONLINE APPENDIX FIGURE III: Distribution of Math Test Scores in 3rd Grade





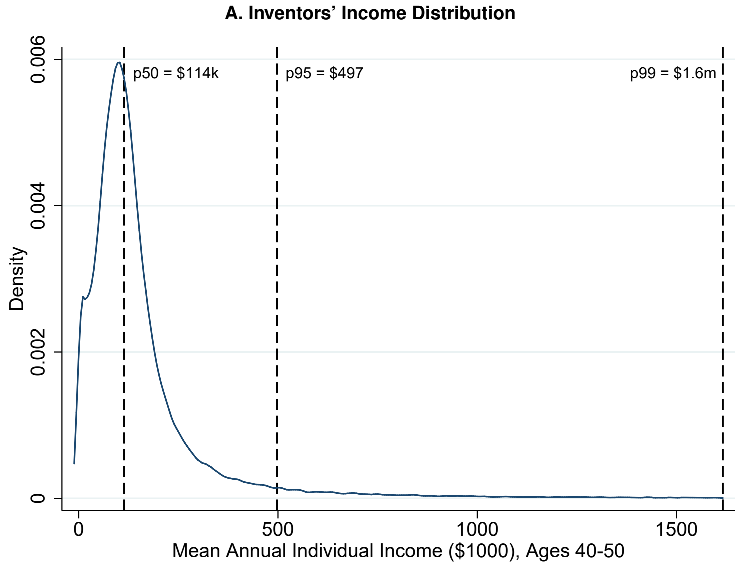
Notes: These figures present kernel densities of 3rd grade math test scores for children in the 1979-1985 birth cohorts who attended New York City public schools. Test scores, which are based on standardized tests administered at the district level, are normalized to have mean zero and standard deviation one by year and grade. In Panel A, we divide children into two groups based on whether their parents' incomes fall below the 80th percentile of the income distribution of parents' income in the New York City schools sample. Panel B compares boys and girls.

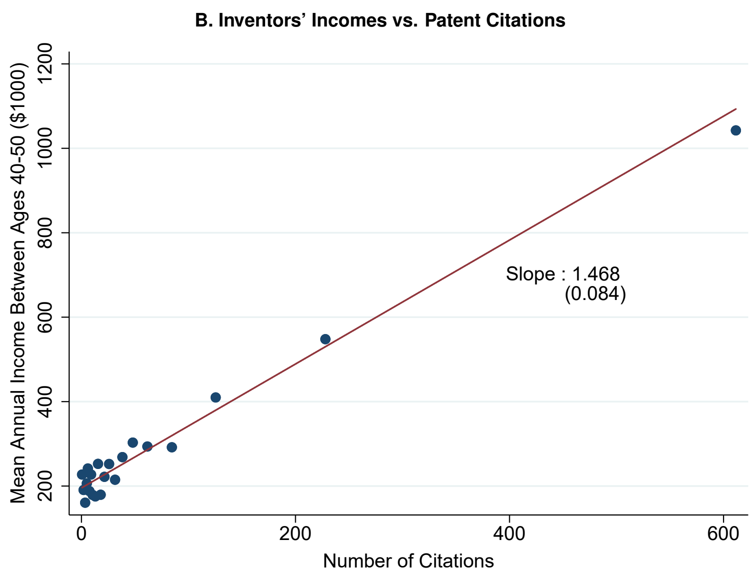
ONLINE APPENDIX FIGURE IV: Female Inventor Share and Gender Stereotype Adherence



Notes: This figure plots the share of inventors who are female vs. Pope and Syndor's (2010) gender stereotype adherence index by state. Female inventor shares are taken directly from Figure Xa; see notes to that figure for details. The stereotype adherence index is computed as $S = (N_{m,\text{math\&science}}/N_{f,\text{math\&science}} + N_{f,\text{reading}}/N_{m,\text{reading}})/2$, where $N_{g,s}$ denotes the number of students of gender $g \in \{m, f\}$ who score among the top 5% of students in their state in subject $s \in \{\text{math \& science}, \text{reading}\}$ in 8th grade. The index S measures the degree to which students adhere to the typical gender stereotype that boys do better at math/science and girls do better in reading; higher values represent greater adherence to this stereotype. The solid best-fit line is estimated using an unweighted OLS regression (slope = -4.49, standard error = 1.42).

ONLINE APPENDIX FIGURE V: Income of Inventors





Notes: Panel A plots a kernel density of the distribution of inventors' income, measured as mean annual income over ages 40-50 in 2012 dollars. Income is measured at the individual level and includes both labor and capital income. For scaling purposes, the top and bottom percentiles of the distribution are omitted. The dashed lines mark the median, 95th percentile, and 99th percentile of the distribution. In both panels, the sample consists of all individuals in our full inventors sample born between the ages of 1959-1962, for whom we see income at all ages between 40 and 50. Panel B presents a binned scatter plot of average annual income between ages 40 and 50 vs. the total number of citations an inventor obtains. For this panel, we further limit the sample to the 13,875 individuals who applied for a patent in 1996 to maximize the time horizon over which we can measure future citations. This plot is constructed by dividing citations into 21 bins and plotting mean income vs. mean citations within each bin. The first 19 bins include inventors in the first 19 ventiles (5% bins) of the citations distribution, while the last two bins plot the same relation for the 95th to 98th percentiles and the 99th percentile of the citation distribution. The best fit line and slope shown on the figure are estimated using an OLS regression on the 21 points, weighted by the number of inventors in each bin. The standard error of the slope estimate is reported in parentheses.