Who Becomes an Inventor? The Importance of Exposure to Innovation

Codebook for Online Data Tables

November 2017

Table 1: Origins of Inventors

Innovation Rates by Childhood CZ/State, Gender and Parent Income

Tables 1a and 1b show patenting outcomes for children born in 1980-1984 by the commuting zone (Table 1a) or state (Table 1b) in which they grew up, gender, and parent income. We restrict the sample to U.S. citizens as of 2013 to exclude individuals who are likely to have immigrated to the U.S. as adults, for whom we cannot measure parent income.

We define a child as an inventor if he or she is listed on a patent application between 2001 and 2012 or grant between 1996 and 2014 (see Section II.B of the paper), and as a highly-cited inventor if he or she is among the 5% of inventors with the most patent citations by 2014 within his birth cohort. Each child is assigned a commuting zone (CZ) or state based on ZIP code from which his or her parents filed their 1040 tax return in the year the child was first claimed as a dependent (which is typically 1996, as our tax data begin in 1996). Parents are assigned percentile ranks by ranking them based on their mean household income from 1996 to 2000 relative to other parents with children in the same birth cohort. See Section II of the paper for further details on the sample construction and variable definitions.

We also report the share of inventors and highly-cited inventors broken down by patent category. We classify patents into technology categories using the classification developed in the NBER Patent Data Project (Hall et al. 2001). We assign each inventor to the category in which he or she patents most often in our sample frame, breaking ties randomly.

We provide statistics on the fraction of inventors by childhood CZ or state pooling all children, by gender, and by parent income quintile.

There is one row in each of these tables per CZ or state. Cells with less than 250 children are omitted.

Users interested in correlating these measures with other CZ characteristics can download a set of CZ-level characteristics from Chetty et al. (2014, Online Data Table 8 [xls] [stata] [codebook]).

References:

Hall, B., A. Jae, and M. Trajtenberg. "The NBER Patent Citation Data File: Lessons, Insights and Methodological Tools." *NBER Working Paper No. 8498*, 2001.

Chetty, R., N. Hendren, P. Kline, and E. Saez. "Where is the Land of Opportunity? The Geography of Intergenerational Mobility in the United States." *Quarterly Journal of Economics* 129(4): 1553-1623, 2014

Codebook for Table 1a – Innovation Rates by Childhood CZ

Variable	Description
par_cz	Childhood commuting zone of residence
par_czname	Commuting zone name
par_state	Childhood state Federal Information Processing Standard (FIPS) code; CZs that cross state borders are assigned to the state which contains the largest population in the CZ, based on the 2010 Census
par_stateabbrv	Two-letter state abbreviation
kid_count	Number of children
inventor	Share of children who go on to become inventors
top5cit	Share of children with patent citations in top 5% of their birth cohort, using total number of citations
inventor_cat_[c]	Share of children who patent in technology category [c]. Technology categories [c] are: 1 – Chemical 2 – Computers and Communications 3 – Drugs and Medical 4 – Electrical and Electronic 5 – Mechanical 6 – Others 7 – Design and Plant
top5cit_cat_[c]	Share of children who patent in technology category [c] and have total patent citations in top 5% of their birth cohort
[outcome]_g_m	Identical to variable [outcome], but restricting the sample to males.
[outcome]_g_f	Identical to variable [outcome], but restricting the sample to females.
[outcome]_pq_[quintile]	Identical to variable [outcome], but restricting the sample to children whose parental income is in quintile [quintile] of the parent income distribution of the children's birth cohort.

Codebook for Table 1b – Innovation Rates by Childhood State

Table 1b contains the same variables as Table 1a with the exception of par_cz and par_czname, since statistics are reported at the state level. Variables in Table 1b are defined identically to variables in Table 1a, except that all statistics are computed directly at the state level in the microdata.

Table 2: Careers of Inventors

Innovation Rates by Current CZ/State, Gender, Year of Birth, and Age

Tables 2a and 2b report patenting outcomes for individuals aged 20 to 80 in years 1996-2012 by year of birth, gender, age and commuting zone (CZ) of residence (Table 2a) or state (Table 2b).

We report the fraction of individuals who file a patent application in a given year as well as the fraction of individuals who file a patent application in that year that is subsequently granted. We observe patent applications in years 2001 to 2012 and patent grants in years 1996 to 2014 (see Section II.B of the paper). All patent grants are public. For a fee, applicants can choose to have their filing kept secret; 15% of applicants choose to do so, and these patent applications do not appear in our data.

Since the grant data span 1996-2014, the data on patents subsequently granted are censored at the beginning and end of the sample frame; for instance, the grantee variable for 1996 includes only patent applications that were filed and granted in 1996, while the grantee variable for 2012 includes only patent applications filed in 2012 that were granted by 2014. This censoring leads to lower grantee rates at the beginning and end of our sample window.

We also report the average number of patent grants by patent category. We classify patents into technology categories using the classification developed in the NBER Patent Data Project (Hall et al. 2001). Note that an inventor may patent in more than one category in a given year.

The population counts reported for each cell are computed from a 10% sample of the taxpayer database.

There is one row in each of these tables for each CZ/state, year of birth, and age. Cells with fewer than 250 observations in the population (i.e., 100% sample) are omitted.

Users interested in correlating these measures with other CZ characteristics can download a set of CZ-level characteristics from Chetty et al. (2014, Online Data Table 8 [xls] [stata] [codebook]).

References:

Hall, B., A. Jae, and M. Trajtenberg. “The NBER Patent Citation Data File: Lessons, Insights and Methodological Tools.” *NBER Working Paper No. 8498*, 2001.

Chetty, R., N. Hendren, P. Kline, and E. Saez. “Where is the Land of Opportunity? The Geography of Intergenerational Mobility in the United States.” *Quarterly Journal of Economics* 129(4): 1553-1623, 2014

Codebook for Table 2a – Innovation by Current CZ, Year of Birth, and Age

Variable	Description
cz	Current commuting zone of residence
czname	Commuting zone name
state	Current state Federal Information Processing Standard (FIPS) code; CZs that cross state borders are assigned to the state which contains the largest population in the CZ, based on the 2010 Census
stateabbrv	Two-letter state abbreviation
cohort	Year of birth
age	Age at which patenting outcomes are measured
year	Calendar year
count	Number of individuals in population
applicant	Fraction of individuals who apply for a patent in current calendar year
grantee	Fraction of individuals who apply for a patent in current year that Is subsequently granted
num_grants	Average number of patents grants per individual, by application year
grantee_cat_[c]	Fraction of individuals granted a patent in technology category [c], by application year. Technology categories [c] are: 1 – Chemical 2 – Computers and Communications 3 – Drugs and Medical 4 – Electrical and Electronic 5 – Mechanical 6 – Others 7 – Design and Plant
[outcome]_g_m	Identical to variable [outcome], but restricting the sample to males.
[outcome]_g_f	Identical to variable [outcome], but restricting the sample to females.

Codebook for Table 2b – Innovation by Current State, Year of Birth, and Age

Table 2b contains the same variables as Table 2a with the exception of cz and czname, since statistics are reported at the state level. Variables in Table 2b are defined identically to variables in Table 2a, except that all statistics are computed directly at the state level in the microdata.

Table 3: Innovation Rates by College

This table presents estimates of students' patent rates by the college they attended. We define the college each child attends as the institution the child attended for the greatest amount of time during the four calendar years in which the child turned 19-22.

The sample includes all children born in the 1980-84 birth cohorts who attend college between the ages of 19-22. We restrict the sample to U.S. citizens as of 2013 to exclude individuals who are likely to have immigrated to the U.S. as adults (for whom we cannot measure parent income).

We define an individual as an inventor if he or she is listed on a patent application between 2001 and 2012 or grant between 1996 and 2014 (see Section II.B of the paper), and as a highly-cited inventor if he or she is among the 5% of inventors with the most patent citations by 2014 within his or her birth cohort.

In addition to patenting outcomes by institution for all students, we provide outcomes by students' parent income quintile. Parents are assigned percentile ranks by ranking them based on their mean household income from 1996 to 2000 relative to other parents with children in the same birth cohort.

For each college, we report the share of students who are inventors (unconditional and conditional on parents' income quantile), the share of students in the top 5% of the patent citation distribution of their birth cohort (among all inventors matched to a college), as well as the total number of patents granted to students and patent citations received by students.

Following established disclosure standards, we report estimates for each college using regression models that pool data across several colleges. As described in Chetty et al. (2017, Appendix C), the degree of error due to this blurring procedure is smaller than the degree of sampling error in the estimates.

There is one row in this table for each college. Colleges with less than 10 inventors are omitted.

Users interested in correlating these measures with other college-level characteristics can download a set of college-level statistics from Chetty et al. (2017), Online Data Table 2 ([xls] [stata] [codebook]) and Online Data Table 10 ([xls] [stata] [codebook]).

Reference:

Raj Chetty, John Friedman, Emmanuel Saez, Nicholas Turner, and Danny Yagan. "Mobility Report Cards: The Role of Colleges in Intergenerational Mobility." *National Bureau of Economic Research Working Paper No. 23618*, July 2017.

Codebook for Table 3 – Innovation Rates by College

Variable	**Description**
super_opeid	Institution OPEID / Cluster ID when combining multiple OPEIDs
instnm	Name of Institution / Super-OPEID Cluster
count	Number of students
count_pq_[quintile]	Number of students with parents in quintile [quintile] of the income distribution
inventor	Share of inventors among students
inventor_pq_[quintile]	Share of inventors among students with parents in quintile [quintile] of the income distribution
top5cit	Share of individuals with total patent citations in top 5% of their birth cohort among all inventors matched to a college
total_patents	Total number of patents granted to students
total_cites	Total number of patent citations obtained by students

Table 4a: Income Distributions of Inventors by Year and Age

This table presents key statistics on the distribution of inventors' income by calendar year and age for years 1999-2012 and ages 25-70.

We define an individual as an inventor if he or she is listed on a patent application between 2001 and 2012 or grant between 1996 and 2014 (see Section II.B of the paper).

We report statistics on three measures of income: total income, wage earnings, and non-wage income. Wage earnings is defined as the sum of earnings across all W-2 forms received by an individual in a given year. Non-wage income consists of self-employment income and capital income. Total income is the sum of wage earnings and non-wage income. Income is measured prior to the deduction of individual income taxes and employee-level payroll taxes. We measure all monetary variables in 2012 dollars, adjusting for inflation using the consumer price index (CPI-U). We round monetary values to the nearest $100.

There is one row in this table for each calendar year and age. Cells with less than 250 inventors are omitted.

Codebook for Table 4a – Income Distributions of Inventors by Year and Age

Variable	Description
year	Calendar year
age	Age
cohort	Year of birth
count	Number of inventors
total_inc_[stat]	Statistic [stat] of the distribution of inventors' total individual income [stat] is either: p[p] – percentile[p], for [p]= 10, 20, 30, 40, 50, 60, 70, 80, 90, 99 mean - mean
w2_inc_[stat]	Statistic [stat] of the distribution of inventors' W-2 wage earnings
nw_inc_[stat]	Statistic [stat] of the distribution of inventors' non-wage individual income

Table 4b: Income Distributions of Highly-Cited Inventors by Age

This table presents key statistics on the distribution highly-cited inventors' income by age over years 1999-2012 for ages 25-70.

We define an individual as an inventor if he or she is listed on a patent application between 2001 and 2012 or grant between 1996 and 2014 (see Section II.B of the paper), and as a highly-cited inventor if he or she is among the 5% of inventors with the most patent citations by 2014 within his or her birth cohort.

We report statistics on three measures of income: total income, wage earnings, and non-wage income. Wage earnings is defined as the sum of earnings across all W-2 forms received by an individual in a given year. Non-wage income consists of self-employment income and capital income. Total income is the sum of wage earnings and non-wage income. Income is measured prior to the deduction of individual income taxes and employee-level payroll taxes. We measure all monetary variables in 2012 dollars, adjusting for inflation using the consumer price index (CPI-U). We round monetary values to the nearest $100.

There is one row in this table for each age. Cells with less than 250 highly-cited inventors are omitted.

Codebook for Table 4b – Income Distributions of Highly-Cited Inventors by Age

Variable	**Description**
age	Age
count	Number of inventors
total_inc_top5cit_[stat]	Statistic [stat] of the distribution of inventors' total individual income [stat] is either: p[p] – percentile[p], for [p]= 10, 20, 30, 40, 50, 60, 70, 80, 90, 99 mean - mean
w2_inc_ top5cit_[stat]	Statistic [stat] of the distribution of inventors' W-2 wage earnings
nw_inc_ top5cit_[stat]	Statistic [stat] of the distribution of inventors' non-wage individual income