DATA APPENDIX

SMART AND ILLICIT: WHO BECOMES AN ENTREPRENEUR AND DO THEY EARN MORE?

ROSS LEVINE AND YONA RUBINSTEIN

July 2016

I: NLSY79

I.A. NLSY79 Variables

Earnings:	
Annual Hours Worked	NUMBER OF HOURS WORKED IN PAST CALENDAR YEAR R24456.00
Full-time, Full-Year	If the respondent (1) works 50 or more weeks per year and (2) works 2000 or more hours per year, and (3) works 40 or more hours per week, then Full-time, Full-year is set equal to one, otherwise it is set equal to zero.
Earnings	Wages plus income from business. Deflated by the CPI corresponding to when those earnings were realized. Earnings are in 2010 prices.
Demographics and Family:	
Age	The age of the respondent. Current year – year of birth.
College Graduate (or more)	Graduated from college or obtained an advanced degree.
Educational attainment (six categories)	The six educational attainment categories: (i) high school dropouts: less than 12 years of schooling (ii) GED degree (iii) high school graduates: 12 years of schooling (iv) had some college education: 13-15 years of schooling (i) college education: 16 years of schooling (vi) advanced studies: 17+ years of schooling. These are measured at the end of the respondent's educational experience, so that they do not vary over time for a respondent.
Family Income in 1979	The income of the respondent's family in 1979 in 2010-year prices. In those cases where 1979 is missing, we use the earliest year between 1980 and 1981 with a non-missing value. See section I.E. NLSY79 Family Income in 1979 for details.
Father's Education	Years of schooling of the respondent's father. See Section I.D. NLSY79 Imputation of Mother and Father Education for details.
Female	Equals one if the respondent reports being female and zero otherwise.
Mother's Education	Years of schooling of the respondent's mother. See Section I.D. NLSY79 Imputation of Mother and Father Education for details.
Potential experience	Equals the age of the respondent minus the years of schooling minus six, or, if this computation is less than zero, then potential experience set equal to zero.

Two parent family (14)	Equals one if the respondent lived in a two-parent family at the age of 14.
White	Equals one if the respondent reports being white and zero otherwise.
Year of birth	The calendar year in which the respondent was born.
Years of Schooling	The respondent's maximum number of years of schooling, so it does not vary over time for a respondent
Traits, etc.	
AFQT	Armed Forces Qualifications Test score measures the aptitude and trainability of the respondent. Collected during the 1980 NLSY79 survey, the AFQT score is based on arithmetic reasoning, world knowledge, paragraph comprehension, and numerical operations. It is frequently employed as a general indicator of cognitive skills. This AFQT score is measured as a percentile of the NLSY79 survey, with a median value of 50.
Applied for Patent (residual standardized)	We set Applied for Patent equal one if the respondent in 2010 answered, "yes" to the question, "Has anyone, including yourself, ever applied for a patent for work that you significantly contributed to?" Since Applied for Patent is obtained decades after a person becomes prime age, we collect the residuals from a regression of Applied for Patent on education, AFQT, Rosenberg Self-Esteem, Rotter Locus of Control, the Illicit Index, and year of birth. We then standardize these residuals to obtain Applied for Patent (residual standardized), which has a mean of zero and a standard deviation of one.
Entrepreneur (residual standardized)	We set the variable Entrepreneur equals one if the respondent in 2010 answers, "yes" to the question, "Do you consider yourself to be an entrepreneur?" In posing the question, the NLSY79 defines an entrepreneur as "someone who launches a business enterprise, usually with considerable initiative and risk." Since Entrepreneur is obtained decades after a person becomes prime age, we collect the residuals from a regression of Entrepreneur on education, AFQT, Rosenberg Self-Esteem, Rotter Locus of Control, the Illicit Index, and year of birth. We then standardize these residuals to obtain Entrepreneur (residual standardized), which has a mean of zero and a standard deviation of one.
Force (raw)	This equals one if the respondent reports in the 1980 survey ever using force to obtain something.
Illicit Activity Index (standardized)	This is constructed based on the answers to 20 questions in the 1980 survey, where 17 are questions about "delinquency" and 3 are about run ins with the "police." The delinquency questions cover issues associated with damaging property, fighting at school, shoplifting, robbery, using force to obtain things, assault, threatening to assault somebody, drug use, dealing drugs, gambling, etc. The "police" questions involve being stopped by the policy, charged with an illegal activity, or convicted, all for activities other

	than minor traffic offenses. For each question, we assign the value one if the person engaged in that activity and zero otherwise. For each respondent, we then add these values and divide by 20. We then standardize the values by subtracting the sample mean and dividing by the standard deviation, so that the Illicit Activity Index has a mean of zero and a standard deviation of one. We provide a more detailed explanation in Section I.B. NLSY79 Illicit Activity Index.
Rosenberg Self-Esteem (standardized)	Rosenberg Self-Esteem score is based on a ten-part questionnaire given to all NLSY79 participants in 1980. It measures the degree of approval or disapproval of one's self. The values range from six to 30, where higher values signify greater self-approval. Rosenberg Self-Esteem (standardized) standardizes the score, so that it has a mean of zero and a standard deviation of one.
Rotter Locus of Control (standardized)	Rotter Locus of Control measures the degree to which respondents believe they have internal control of their lives through self-determination relative to the degree that external factors, such as chance, fate, and luck, shape their lives. It was collected as part of a psychometric test in the 1979 NLSY79 survey. The Rotter Locus of Control ranges from 4 to 16, where higher values signify less internal control and more external control. This is standardized, so that it has a mean of zero and a standard deviation of one.
Steal 50 or less (raw)	This equals one if the respondent reports in the 1980 survey stealing something worth $50 or less during the year.
Stopped by Police (raw)	This equals one if the respondent reports in the 1980 survey ever being stopped by the police.
Employment type	
Salaried	From the NLSY79's unified class of worker (R24455.10), there are four responses for working respondents: (1) Private company, including non-profit, (2) government, (3) self-employed, and (4) those working without pay, including in family businesses. We set Salaried equal to one if the respondent's class of worker is either "(1)" or "(2)"and zero otherwise.
Self-employed	From the NLSY79's unified class of worker (R24455.10), there are four responses for working respondents: (1) Private company, including non-profit, (2) government, (3) self-employed, and (4) those working without pay, including in family businesses. We set Self-employed equal to one if the respondent's class of worker is "(3)" and zero otherwise.
Incorporated Self-employed	If a respondent is self-employed, the NLSY79 further asks whether the business is incorporated or not. If the respondent is self-employed and the business is incorporated, then Incorporated Self-employed equals one and it is zero otherwise. See Section I.C. NLSY79 Incorporated Self-Employment Coding Details.
Unincorporated Self-employed	If a respondent is self-employed, the NLSY79 further asks whether the business is incorporated or not. If the respondent is self-

	employed and the business is unincorporated, then Unincorporated Self-employed equals one and it is zero otherwise.

I.B. NLSY79 Illicit Activity Index

In this subsection, we first describe the core data from the NLSY79 survey and then provide details on the construction of the index

I.B.1. The Core Data

The Illicit Activity index is based on questions from the 1980 survey. We use two types of questions on illicit activities "delinquency" and "police" questions.

We use data on 17 of the 20 questions on "delinquency" provided by the NLSY79. We do not use the other three questions that were only posed to people who were 17 years old or younger in 1979. Thus, these questions were only asked of about 30% of the sample (3,898 out of 12,686). Including these variables would reduce the sample by about 70%. The survey asks about whether—and how many times—the respondent engaged in the delinquent act. For example, one of the questions asks how many times the respondent smoked marijuana/hashish in the past year. We list all 20 questions below and indicate which ones we use in constructing the Illicit Activity Index. (This is titled "Table on the distribution of responses to questions on delinquency and police.")

Use two versions of the responses to the delinquency questions:

(1) There is the "0-6 intensive" version that uses the actual number of times the respondent engaged in the delinquent act, where there are seven answers categorized from zero to six; and

(2) There is our primary, core ("extensive") version that uses the values of zero or one in coding the responses to the delinquency questions, i.e., we code the values as either the person did or did not engage in the act at least once.

With respect to the "police" questions, we use the three questions on interactions with the police provided by the NLSY79. These three questions offer zero/one options for responses. For example, one of the questions asks, were you "… ever convicted on illegal activity charges other than minor traffic offense?" We list these questions below also.

Here is a listing of the core data:

NLSY79 Reference Number	Question Name	Definition	Year of survey	Sample
Delinquency questions used in constructing the Illicit Activity Index				
R03049.00	DELIN-4	ILLEGAL ACTIVITY 80 INT - TIMES INTENTIONALLY DAMAGED PROPERTY IN PAST YEAR	1980	ALL
R03050.00	DELIN-5	ILLEGAL ACTIVITY 80 INT - TIMES FOUGHT AT SCHOOL OR WORK IN PAST YEAR	1980	ALL
R03051.00	DELIN-6	ILLEGAL ACTIVITY 80 INT - TIMES SHOPLIFTED IN PAST YEAR	1980	ALL
R03052.00	DELIN-7	ILLEGAL ACTIVITY 80 INT - TIMES STOLEN OTHER'S BELONGINGS PAST YR (WORTH <$50)	1980	ALL
R03053.00	DELIN-8	ILLEGAL ACTIVITY 80 INT - TIMES STOLEN OTHER'S BELONGINGS PAST YR (WORTH >$50)	1980	ALL
R03054.00	DELIN-9	ILLEGAL ACTIVITY 80 INT - TIMES USED FORCE TO OBTAIN THINGS IN PAST YEAR	1980	ALL
R03055.00	DELIN-10	ILLEGAL ACTIVITY - TIMES SERIOUSLY THREATENED TO HIT/HIT SOMEONE PAST YEAR	1980	ALL
R03056.00	DELIN-11	ILLEGAL ACTIVITY 80 INT - TIMES ATTACKED W/INTENT TO INJURE/KILL IN PAST YEAR	1980	ALL
R03057.00	DELIN-12	ILLEGAL ACTIVITY 80 INT - TIMES SMOKED MARIJUANA/HASHISH IN PAST YEAR	1980	ALL
R03058.00	DELIN-13	ILLEGAL ACTIVITY - TIMES USED OTHER DRUGS/CHEMICALS TO GET HIGH IN PAST YEAR	1980	ALL
R03059.00	DELIN-14	ILLEGAL ACTIVITY 80 INT - TIMES SOLD MARIJUANA/HASHISH IN PAST YEAR	1980	ALL
R03060.00	DELIN-15	ILLEGAL ACTIVITY 80 INT - TIMES SOLD HARD DRUGS IN PAST YEAR	1980	ALL
R03061.00	DELIN-16	ILLEGAL ACTIVITY 80 INT - TIMES ATTEMPTED TO "CON" SOMEONE IN PAST YEAR	1980	ALL
R03062.00	DELIN-17	ILLEGAL ACTIVITY 80 INT - TIMES TAKEN AUTO W/OUT OWNER PERMISSION IN PAST YEAR	1980	ALL
R03063.00	DELIN-18	ILLEGAL ACTIVITY 80 INT - TIMES BROKEN INTO A BUILDING IN PAST YEAR	1980	ALL
R03064.00	DELIN-19	ILLEGAL ACTIVITY 80 INT - TIMES KNOWINGLY SOLD/HELD STOLEN GOODS IN PAST YEAR	1980	ALL
R03065.00	DELIN-20	ILLEGAL ACTIVITY 80 INT - TIMES AIDED IN GAMBLING OPERATION IN PAST YEAR	1980	ALL
Police questions (zero/one answers)				
R03067.00	POLICE-1	EVER "STOPPED" BY POLICE FOR OTHER THAN MINOR TRAFFIC OFFENSE?	1980	ALL
R03071.00	POLICE-2	EVER CHARGED WITH ILLEGAL ACTIVITY? 80 INT (EXC MINOR TRAFFIC OFFENSE)	1980	ALL

R03078.00	POLICE-3	EVER CONVICTED ON ILLEGAL ACTIVITY CHARGES OTHER THAN MINOR TRAFFIC OFFENSE?	1980	ALL

Delinquency questions NOT used in constructing the Illicit Activity Index

NLSY79 Reference Number	Question Name	Definition	Year of survey	Sample
R03046.00	DELIN-1	ILLEGAL ACTIVITY 80 INT - TIMES RUN AWAY FROM HOME IN PAST YR (AGE 17 OR UNDER)	1980	R<=17
R03047.00	DELIN-2	ILLEGAL ACTIVITY 80 INT - TIMES SKIPPED SCHOOL DAY IN PAST YR (AGE 17 OR UNDER)	1980	R<=17
R03048.00	DELIN-3	ILLEGAL ACTIVITY - TIMES DRANK ALCOHOLIC BEVERAGES PAST YR (AGE 17 OR UNDER)	1980	R<=17

These last three questions are excluded because they were only asked of respondents who were 17 years or younger in 1979 (born between 1962 and 1964). Including them would reduce the sample by 70%.

I.B.2. Constructing the Index

The Illicit Index

1. Sample: of the 12,686 individuals in the NLSY79 survey, 1,310 have missing data on one of the 17 delinquency questions that we include in our Index. Of the remaining 11,376 individuals, 19 have missing data on one of the three police questions.

This is tabulated as follows:

Variable / Selection Criteria	**Persons**	**Dropped**
Initial number of persons	12,686	0
Missing one of the 17 questions	11,376	1,310
Missing one of the 3 questions	11,357	19
Final Number of Person-Observations	**11,357**	

2. We use the (i) responses to the three police questions (which offer zero/one responses in the NLSY79 survey) and (ii) the "one/zero" responses to the 17 delinquency questions, i.e., we use the extensive versions of the answers to the delinquency questions.
3. For each respondent, we sum the responses to the twenty zero/one responses and divide by 20. We then standardize the values by subtracting the sample mean and dividing by the standard deviation, so that the Illicit Activity Index has a mean of zero and a standard deviation of one.

The Illicit Activity Index 0-6 Intensive

1. We use the full answers to the 17 delinquency variables concerning the number of times the respondent engaged in the activity. Specifically, there are seven possible answers: (0) never, (1) once, (2) twice, (3) 3-5 times, (4) 6-10 times, (5) 11-50 times, and (6) more than 50 times.
2. For the Illicit Activity Index 0-6 Intensive, we assign the values 0, 1, 2, 4, 8, 30, and 50 to the seven answers.
3. We then (a) compute the standardized value of each of the 20 questions ((value – mean)/standard deviation)) and then (b) sum the values and divide by 20.

The two, Illicit Activity Index (standardized) and the Illicit Activity Index 0-6 intensive (standardized), are very highly correlated (0.91) and the paper's results hold when using either Index. We illustrate this robustness in the Appendix Tables: APPENDIX TABLE VIIA, APPENDIX TABLE VIII, and APPENDIX TABLE XI.

The Distribution of Responses to Questions on Delinquency and Police:

Question	N	Min	Median	Mean	Max	NEVER	ONCE	TWICE	TIMES_3_5	TIMES_1_10	TIMES_11_50	TIMES_5
Delinquency questions used in constructing the Illicit Activity Index												
DAMAGED	11734	0	0	0.36	6	0.82	0.09	0.04	0.04	0.01	0.00	0.00
FOUGHT	11800	0	0	0.56	6	0.72	0.13	0.06	0.06	0.02	0.01	0.00
SHOPLIFTED	11788	0	0	0.53	6	0.74	0.12	0.05	0.05	0.02	0.01	0.00
STEEL50M	11788	0	0	0.37	6	0.81	0.09	0.04	0.04	0.01	0.01	0.00
STEEL50P	11776	0	0	0.11	6	0.94	0.03	0.01	0.01	0.00	0.00	0.00
FORCE	11794	0	0	0.10	6	0.95	0.03	0.01	0.01	0.00	0.00	0.00
THREAT	11785	0	0	0.82	6	0.63	0.16	0.08	0.08	0.03	0.02	0.01
ATTACK	11792	0	0	0.20	6	0.89	0.06	0.02	0.02	0.01	0.00	0.00
MARIJUANA	11722	0	0	1.84	6	0.53	0.09	0.04	0.07	0.05	0.07	0.15
DRUGES	11698	0	0	0.60	6	0.81	0.05	0.03	0.04	0.03	0.03	0.02
MARIJUANA_SOLD	11693	0	0	0.33	6	0.89	0.03	0.02	0.02	0.02	0.01	0.01
DRUGES_SOLD	11717	0	0	0.07	6	0.97	0.01	0.00	0.00	0.00	0.00	0.00
CON	11721	0	0	0.48	6	0.78	0.09	0.05	0.05	0.02	0.01	0.01
AUTO	11752	0	0	0.14	6	0.92	0.04	0.01	0.01	0.00	0.00	0.00
BROKEN	11748	0	0	0.12	6	0.94	0.03	0.01	0.01	0.00	0.00	0.00
SOLD	11749	0	0	0.23	6	0.89	0.06	0.02	0.02	0.01	0.00	0.00
GAMBELING	11737	0	0	0.06	6	0.98	0.01	0.00	0.00	0.00	0.00	0.00
Police questions (zero/one answers) used in constructing the Illicit Activity Index												
STOPPED_POLICE	12129	0	0	0.19	1	0.81	0.19	0.00	0.00	0.00	0.00	0.00
CHARGED	12136	0	0	0.11	1	0.89	0.11	0.00	0.00	0.00	0.00	0.00
CONVICTED	12130	0	0	0.06	1	0.94	0.06	0.00	0.00	0.00	0.00	0.00
AVG						**0.84**	**0.07**	**0.03**	**0.03**	**0.01**	**0.01**	**0.01**

I.C. NLSY79 Incorporated Self-employment Coding Details

We follow the following process for coding incorporated self-employment.

1. The NLSY79 provides information on (a) the class of worker, including whether the respondent (R) is salaried or self-employed, and (b) whether R's business is incorporated.

2. From 1994 onward, the NLSY79 notes that whenever R's job is the same as the job in the last interview, class of worker and incorporation status are only reported if the information has changed. It is coded as missing if there has been no change since the last interview.

3. For example, if person A has been continuously self-employed by "AL Consulting, Inc." for several years, A's "raw" data might look like this:

Year	COW(job#1)	INCORP(job#1)
2000	4 (SE)	1 (yes)
2002	-4	-4
2004	-4	-4
2006	-4	-4
2008	-4	-4
2010	-4	-4

Even though job#1 refers to the same job (AL Consulting, Inc.) in each of these interviews, COW and INCORP are missing after the first year because they are not re-asked.

4. The NLSY79 solves this problem for class of worker. They appropriately "fill in" the information on class of worker for each R rather than leaving data entries as "missing," e.g., see the class of worker variable for Job #1 in the NLSY79 Navigator). The COWALL variables (e.g., R4587905 = COWALL-EMP1_1994, which is COW for job#1 in 1994) have been "filled in" to carry the old information forward.
5. The NLSY79, however, did not "fill in" the incorporation status. The INCORP variables (e.g., R4587000=QES1-56E_1994, which is INCORP for job#1 in 1994) have not. So if one is pairing (created) COWALL variables with (raw) INCORP variables, the (fake) data will look like this:

Year	COWALL-EMP1	INCORP(job#1)
2000	4 (SE)	1 (yes)
2002	4	-4
2004	4	-4
2006	4	-4
2008	4	-4
2010	4	-4

Clearly, there are many cases where INCORP is missing even though COWALL=SE.

6. There is a straightforward procedure for addressing this coding issue. If R is self-employed, use the incorporation status of the last interview to appropriately "fill in" missing values (i.e., fill in the "valid skips"). After following this procedure, incorporate status has 1.5% missing values based on the sample of individuals in the Table I summary statistics.

7. We make a few additional adjustments for Rs who (a) are self-employed but (b) have missing values of incorporated business status after following the above procedure in a survey year.

   7.a. We find 3 person-year observations after 1994 where R is self-employed and R reports having the same job as last year and R was incorporated last year. We code R as incorporated. (Same Job as Last Year)

   7.b. We find 8 person-year observations (after 1994) in which the NLSY79 indicates in survey t+1 that (a) R is incorporated and (b) has the same job as last year, so we code R as incorporated in survey t. (Same Job as Next Year)

   7.c. We find 7 person-year observations prior to 1994 in which a self-employed person in period t and t-1 has a missing value for incorporated status in survey t, and was incorporated self-employed in survey t-1. We code these Rs as incorporated in survey t. (Pre-1994: Same Job as Last Year)

   7.d. We find 9 person-year observations in which a self-employed person in survey t and t+1 has a missing value for incorporated status in survey t but is incorporated self-employed in survey t+1. We code these Rs as incorporated in survey t. (Pre-1994: Same Job as Next Year)

   7.e. After this, there are 2 additional person-year observations in which the incorporated status is missing, but in which one of Rs other jobs, i.e., Job 2 – Job 5, is incorporated. We code these two observations as incorporated. Across Job Categories)

   7.f. The results are robust to keeping these 29 person-year observations as missing, as shown in Appendix Tables. See APPENDIX TABLE I, APPENDIX TABLE II, APPENDIX TABLE IV, and APPENDIX TABLE VIIB.

   Specifically, from the STATA program for the 1,936 person-year observations coded as incorporated self-employed, we tabulate the following:

Incorporated			
Source	**Freq.**	**%**	**Cum.**
NLSY79 raw data	1,501	77.53	77.53
NLSY79 post 1994 procedure	4,06	20.97	98.50
Same job as last Year	3	0.15	98.66
Same job as next year	8	0.41	99.07
Pre-1994: same job as last year	7	0.36	99.43
Pre-1994: same job as next year	9	0.46	99.90
Across job categories	2	0.10	100.0

I.D. NLSY79: Imputation of Mother and Father Education

Of the 132,681 person-year observation (10,719 individuals) covered in Table I, 125,291 (10,093) have mother's education and 115,216 (9,263) have father's education.

If the NLSY79 does not report data on the education of the mother or the father, we use the following two imputation procedures.

1. Partner Imputation. If one parent's education is missing, we use the other parent's.

2. Mean Imputation. If both parent's education are missing, we use the mean education of parents, differentiating by race (Black, Hispanic, White) and gender.

As shown in Appendix Tables, the results are robust to excluding these imputed measures.

The following tables detail each imputation procedure (one observation per individual) for the 10,719 individuals in our base sample (see next page).

Mother Education			
Source	**Freq.**	**%**	**Cum.**
NLSY79 raw data	10,093	94.16	94.16
Imputed using father's education	294	2.74	96.90
Imputed using group's mean	332	3.10	100.00

Father Education			
Source	**Freq.**	**%**	**Cum.**
NLSY79 raw data	9,263	86.42	86.42
Imputed using father's education	1,124	10.49	10.49
Imputed using group's mean	332	3.10	3.10

I.E. NLSY79: Family Income in 1979

For Family Income in 1979, we use the non-zero values of the variable R0217900, which is truncated at $75,000 in 2010-year prices. We use the earliest non-missing value in 1979-1981. In 81% of the cases, this is 1979. In 13%, it is 1980; and in 3.3%, it is 1981.

Thus, of the 132,681 person-year observations in Table I, 8,676, we have non-missing data for Family Income in 1979 for 97.3% of those observations

For the remaining 2.7% (3,598 person-year observations), we impute family income by using the mean value of family income by race (Black, Hispanic, White).

I.F. Data Processing NLSY79: Sample Selection Criteria

The table on the next page details how we arrive at the number of observations in each table using NLSY79 data.

Data Processing NLSY79

Variable / Selection Criteria	Year-Person	Individuals	Dropped
Initial number of observations	317,150	12,686	0
Interviewed	243,641	12,686	73,509
Age between 25 to 55	166,250	12,264	77,391
Salaried or Self-Employed	143,583	11,780	22,667
AFQT	137,272	11,133	6,311
Rotter Score	136,037	11,020	1,235
Rosenberg Score	132,681	10,719	3,356
School Years	132,681	10,719	0
Number of Observations	**132,681**	**10,719**	**184,469**
Tables and Figures			
Tables I and II: Demographics, Labor Market Outcomes and Home Environment			
Number of Observations	**132,681**	**10,719**	184,469
Table III: Switching Between Unincorporated and Incorporated Self-Employment			
Table I and	132,681	10,719	184,469
Self-employed, first year in spell	4,118	2,799	128,563
At most one switch within self-employment spell	4,083	2,786	35
Number of Observations	**4,083**	**2,786**	0
Table IV: Job Task Requirements by Employment Type			
Table I and	132,681	10,719	184,469
Excluding missing occupation	131,949	10,674	732
Number of Observations Panel B.1	**131,949**	**10,674**	
Last job as salaried worker	120,156	10,218	12,525[1]
Number of Observations Panel B.2	**120,156**	**10,218**	
Table VII: Selection into Employment Types on Cognitive, Noncognitive and Family Traits			
Table I and	132,681	10,719	
Illicit Index	125,166	10,055	7,515
Number of Observations	**125,166**	**10,055**	

[1] This drop is relative to the observations in Table I.

Table VIII: Differences in Job Task Requirements of Businesses by Individual Traits

Table VII and	125,166	10,055	
Whites	69,503	5,981	55,663
Males	35,012	2,964	34,491
Salaried two years ago	29,754	2,818	5,258
Valid industry codes in year t	29,412	2,817	342
Number of Observations	**29,412**	**2,817**	

Tables IX, X, XI and Figures I and II: Earnings, Levels and First Differences

Table VII and	125,166	10,055	
Whites	69,503	5,981	55,663
Males	35,012	2,964	34,491
Hourly Earnings	32,768	2,924	2,244
Full-Time; Full-Year	23,657	2,595	9,111
Number of Observations	**23,657**	**2,595**	
- First differences regressions	**17,479**	**2,227**	6,178

II: CPS

II.A. Variables

Earnings:	
Annual Hours Worked	Number of hours worked during the past calendar year
Full-time, Full-Year	If the respondent (1) works 50 or more weeks per year and (2) works 2000 or more hours per year, and (3) works 40 or more hours per week, then Full-time, Full-year is set equal to one, otherwise it is set equal to zero.
Earnings	Wages plus income from business. Deflated by the CPI corresponding to when those earnings were realized. Earnings are in 2010 prices.
Demographics:	
Age	The age of the respondent
College Graduate (or more)	Graduated from college or obtained an advanced degree.
Educational attainment (six categories)	The six educational attainment categories are: (i) completed less than 9th grade, (ii) completed between 9th and 11th grade, (iii) graduated from high school, (iv) had some college education, (v) graduated from college, and (vi) obtained an advanced degree.
Female	Equals one if the respondent reports being female and zero otherwise.
Potential experience	Equals the age of the respondent minus the years of schooling minus seven, or, if this computation is less than zero, then potential experience set equal to zero.
White	Equals one if the respondent reports being white and zero otherwise.
Year of birth	The calendar year in which the respondent was born.
Years of Schooling	Total years of educational attainment.
Employment type	
Salaried	The CPS classifies all workers in each year as either salaried or self-employed. Salaried equals one if the respondent is salaried and zero otherwise.
Self-employed	The CPS classifies all workers in each year as either salaried or self-employed. Self-employed equals one if the respondent is salaried and zero otherwise.
Incorporated Self-employed	The CPS classifies all workers in each year as either salaried or self-employed, and among the self-employed, indicates whether individuals are incorporated or unincorporated. Specifically, individuals are asked about their employment class for their main job: "Were you employed by a government, by a private company, a nonprofit organization, or were you self-employed (or working in a family business)?" Those responding that they are self-employed are further asked, "Is this business incorporated?" Incorporated self-employed equals one if the person answers yes, and zero otherwise.

Unincorporated Self-employed	The CPS classifies all workers in each year as either salaried or self-employed, and among the self-employed, indicates whether individuals are incorporated or unincorporated. Specifically, individuals are asked about their employment class for their main job: "Were you employed by a government, by a private company, a nonprofit organization, or were you self-employed (or working in a family business)?" Those responding that they are self-employed are further asked, "Is this business incorporated?" Unincorporated self-employed equals one if the person answers no to this question and yes to being self-employed, and zero otherwise.

II.B. Data Processing CPS: Sample Selection Criteria

The table on the next page details how we arrive at the number of observations in each table using NLSY79 data in the paper.

Data Processing CPS, 1996-2013

Variable / Selection Criteria	**Year-Person**	**Individuals[2]**	**Dropped**
Initial number of observations	3,384,125	--	0
Adult Civilians	2,551,860	--	832,265
Households	2,551,836	--	24
With positive sample weight	2,550,441	--	1,395
Age between 25 to 55	1,500,103	--	1,050,338
Gender, race, education	1,500,103	--	0
Potential experience 0-50	1,500,103	--	0
Valid industry code	1,257,925	--	242,178
Valid occupation code (<997)	1,257,925	--	0
School Years	1,257,925	--	0
Excluding:			
-Farmers and Farm Laborers	1,240,776	--	17,149
-Agriculture	1,226,658	--	14,118
Salaried or self-employed	1,225,886	--	772
Number of Observations	**1,225,886**	**893,780**	**2,126,200**

[2] The number of individuals is the base for our CPS panel data analyses.

Tables

Selection Criteria	Year-Person	Individuals	Dropped
Table I: Demographics and Labor Market Outcomes by Employment Type			
Number of Observations	**1,225,886**	**893,780**	2,126,200
Table IV: Job Task Requirements by Employment Type			
Table I	1,225,886	893,780	
Number of Observations panel A.1	**1,225,886**	**893,780**	
Panel	513,701	257,017	712,185
Salaried worker last year	230,330	230,330	283,371
Number of Observations panel A.2	**230,330**	**230,330**	
Table V: Selection into Unincorporated and Incorporated Self-Employment			
Table IV panel A.2	230,330	230,330	
Number of Observations	**230,330**	**230,330**	
Table VI: Top and Bottom Industries by Nonroutine Job Task Requirements			
Table I	1,225,886	893,780	
Number of Observations	**1,225,886**	**893,780**	

II.C. Matched Sample

We construct a two-year matched panel. The CPS interviews a household for four consecutive months. The next year, the CPS returns to the same location. In most cases, the second interview involves the same household as the first interview.

We follow the guidelines in Madrian and Lefren (2000) for matching CPS households across time. This involves checking the age, race, gender, education, etc. of those interviewed and dropping individuals (for the matched panel sample) where these do not match across the CPS interviews.

We do not find differential selection into the matched-CPS sample once we condition on demographics (and FTFY when conducting the earnings analyses), as shown in the table below.

More specifically, selection into the panel sub-sample is not random. Whites and individuals with larger earnings are more likely to be observed two consecutive years than others. The incorporated and unincorporated self-employed are more likely to be selected into the two-year panel (5.85% and 1.7%) than others. Yet, conditional on standard demographics, such as gender, race, education and potential experience these differences disappear.

A simple (non-parametric) way to observe that differential selection is not a problem when using the matched-CPS sample is to compare the cross-section CPS sample with the Matched-CPS sample by demographic groups. When we restrict the sample to whites, we find much smaller gaps in key measures between the Matched-CPS sample and the cross-section sample. For instance, the gap in years of schooling completed, annual hours worked, and the DOT measures are negligible. The gap in annual earnings between the Matched-CPS sample and the cross-section sample drops by half to 4%. Furthermore, when restricting the sample to FTFY white men, the gap in earnings drops to 1%. The differential selection on earnings into the CPS-Matched sample drops from approximately 7% (more for salaried than incorporated and unincorporated self-employed) to approximately 2%.

For a comparison of the earnings when using the full and matched samples, see Appendix Table IXB: CPS: Earnings Full & Matched Samples, which shows that the results are very similar.

Differences between the Cross-Section CPS and the Matched-Panel CPS

Sample:	Differences in Absolute Terms (All-Matched)					Differences in % (All-Matched)/All				
	All	**Whites**	**White Males**	**White Males FTFY**	**White Males FTFY2K**	**All**	**Whites**	**White Males**	**White Males FTFY**	**Wh Ma FTF**
Panel A: All Types of Workers										
Observations	712185	439193	224825	183224	177032	**58.1%**	52.2%	51.8%	**50.3%**	**5**
Age	-1.4	-1.2	-1.1	-1.0	-1.0	**-3.5%**	-2.9%	-2.7%	**-2.4%**	**-**
White	-0.10	0.00	0.00	0.00	0.00	**-14.4%**	0.0%	0.0%	**0.0%**	
Female	0.00	0.00	0.00	0.00	0.00	**-0.1%**	0.5%			
Years of schooling	-0.2	-0.1	0.0	0.0	0.0	**-1.6%**	-0.4%	-0.3%	**-0.2%**	**-**
Mean earnings	-3853	-2217	-2222	-867	-821	**-8.1%**	-4.3%	-3.5%	**-1.2%**	**-**
Median earnings	-3930	-2225	-2010	-1158	-1171	**-10.9%**	-5.6%	-4.1%	**-2.2%**	**-**
Annual worked hours	-61	-49	-35	3	3	**-3.0%**	-2.4%	-1.6%	**0.1%**	
Full-Time Full-Year	-0.03	-0.03	-0.03	0.00	0.00	**-5.0%**	-4.4%	-3.4%	**-0.1%**	
Nonroutine Analytical	-0.16	-0.06	-0.05	-0.03	-0.03	**-4.0%**	-1.5%	-1.2%	**-0.6%**	**-**
Nonroutine DCP	-0.23	-0.10	-0.09	-0.05	-0.05	**-7.8%**	-3.1%	-2.7%	**-1.5%**	**-**
Nonroutine Manual	0.04	0.01	0.01	0.01	0.01	**3.6%**	1.5%	1.0%	**0.7%**	
Panel B: Salaried										
Observations	647422	392704	196143	160633	155473	**58.4%**	52.3%	52.1%	**50.5%**	**5**
Age	-1.4	-1.2	-1.2	-0.6	-0.7	**-3.6%**	-3.1%	-2.9%	**-1.4%**	**-**
White	-0.10	0.00	0.00	0.00	0.00	**-14.9%**	0.0%	0.0%	**0.0%**	
Female	0.00	0.00	0.00	0.00	0.00	**-0.6%**	0.2%			
Years of schooling	-0.2	-0.1	0.0	0.0	-0.1	**-1.7%**	-0.4%	-0.3%	**-0.3%**	**-**
Mean earnings	-3802	-2236	-2321	126	-124	**-8.2%**	-4.4%	-3.8%	**0.2%**	**-**

Median earnings	-3770	-2067	-2187	0	-405	**-10.4%**	-5.2%	-4.5%	**0.0%**	-
Annual worked hours	-59	-48	-35	-1	-4	**-3.0%**	-2.4%	-1.6%	**0.0%**	-
Full-Time Full-Year	-0.04	-0.03	-0.03	0.00	0.00	**-5.1%**	-4.5%	-3.6%	**0.0%**	
Nonroutine Analytical	-0.16	-0.06	-0.05	-0.02	-0.01	**-4.0%**	-1.5%	-1.2%	**-0.5%**	-
Nonroutine DCP	-0.23	-0.10	-0.09	-0.02	0.05	**-8.0%**	-3.2%	-2.8%	**-0.6%**	
Nonroutine Manual	0.04	0.01	0.01	0.02	0.01	**3.6%**	1.5%	0.8%	**1.5%**	
Panel C: Self-Employed										
Observations	64763	46489	28682	22591	21559	**55.2%**	50.8%	49.8%	**49.0%**	**4**
Age	-0.9	-0.7	-0.6	-0.6	-0.6	**-2.0%**	-1.6%	-1.4%	**-1.3%**	-
White	-0.08	0.00	0.00	0.00	0.00	**-10.2%**	0.0%	0.0%	**0.0%**	
Female	0.01	0.01	0.00	0.00	0.00	**3.2%**	3.1%			
Years of schooling	-0.2	-0.1	-0.1	0.0	0.0	**-1.4%**	-0.4%	-0.4%	**-0.3%**	-
Mean earnings	-3553	-1735	-978	126	150	**-6.1%**	-2.8%	-1.3%	**0.1%**	
Median earnings	-2749	-1627	-1239	0	-821	**-8.0%**	-4.5%	-2.6%	**0.0%**	-
Annual worked hours	-64	-51	-31	-1	-1	**-3.1%**	-2.4%	-1.3%	**0.0%**	
Full-Time Full-Year	-0.03	-0.02	-0.02	0.00	0.00	**-4.3%**	-3.6%	-2.0%	**0.0%**	
Nonroutine Analytical	-0.13	-0.06	-0.04	-0.02	-0.02	**-3.1%**	-1.3%	-0.8%	**-0.4%**	-
Nonroutine DCP	-0.17	-0.08	-0.06	-0.02	-0.02	**-4.5%**	-2.1%	-1.3%	**-0.5%**	-
Nonroutine Manual	0.03	0.01	0.02	0.02	0.02	**3.5%**	1.4%	1.9%	**1.7%**	
Panel D: Self-Employed Unincorporated										
Observations	42785	29506	16440	11804	11101	**56.7%**	51.7%	50.3%	**49.3%**	**4**
Age	-1.0	-0.8	-0.7	-0.7	-0.7	**-2.4%**	-1.9%	-1.7%	**-1.6%**	-
White	-0.09	0.00	0.00	0.00	0.00	**-12.2%**	0.0%	0.0%	**0.0%**	
Female	0.01	0.01	0.00	0.00	0.00	**2.2%**	3.0%			
Years of schooling	-0.2	-0.1	-0.1	-0.1	-0.1	**-1.8%**	-0.6%	-0.6%	**-0.4%**	-
Mean earnings	-3150	-1905	-1375	-210	-124	**-7.7%**	-4.4%	-2.5%	**-0.3%**	-
Median earnings	-2415	-1632	-1026	-102	-405	**-9.8%**	-6.3%	-3.0%	**-0.3%**	-
Annual worked hours	-66	-57	-38	-2	-4	**-3.4%**	-3.0%	-1.8%	**-0.1%**	-
Full-Time Full-Year	-0.03	-0.02	-0.02	0.00	0.00	**-4.4%**	-4.0%	-2.2%	**0.3%**	

Nonroutine Analytical	-0.14	-0.07	-0.04	-0.02	-0.01	**-3.7%**	-1.7%	-1.0%	**-0.5%**	**-**
Nonroutine DCP	-0.13	-0.05	-0.02	0.04	0.05	**-4.1%**	-1.6%	-0.5%	**1.1%**	
Nonroutine Manual	0.03	0.01	0.02	0.01	0.01	**2.7%**	0.6%	1.2%	**0.6%**	
Panel E: Self-Employed Incorporated										
Observations	21978	16983	12242	10787	10458	**52.6%**	49.4%	49.1%	**48.6%**	**4**
Age	-0.6	-0.5	-0.4	-0.4	-0.4	**-1.4%**	-1.1%	-1.0%	**-1.0%**	**-**
White	-0.06	0.00	0.00	0.00	0.00	**-6.7%**	0.0%	0.0%	**0.0%**	
Female	0.01	0.01	0.00	0.00	0.00	**3.1%**	2.1%			
Years of schooling	-0.1	0.0	0.0	0.0	0.0	**-0.4%**	-0.1%	-0.1%	**-0.1%**	**-**
Mean earnings	-1394	-147	467	963	1003	**-1.6%**	-0.2%	0.4%	**0.9%**	
Median earnings	-1390	-609	-592	-310	-728	**-2.5%**	-1.1%	-0.9%	**-0.4%**	**-**
Annual worked hours	-37	-30	-15	2	3	**-1.6%**	-1.3%	-0.6%	**0.1%**	
Full-Time Full-Year	-0.02	-0.02	-0.01	0.00	0.00	**-2.4%**	-2.3%	-1.4%	**-0.2%**	
Nonroutine Analytical	-0.05	-0.02	-0.01	-0.01	-0.01	**-1.1%**	-0.4%	-0.2%	**-0.2%**	**-**
Nonroutine DCP	-0.13	-0.09	-0.08	-0.07	-0.06	**-2.5%**	-1.7%	-1.4%	**-1.2%**	**-**
Nonroutine Manual	0.03	0.02	0.02	0.03	0.03	**3.3%**	2.1%	2.4%	**3.0%**	

III: Job Task Requirements

III.A. Basics

The DOT was first constructed in 1939 to help employment offices match job seekers with job openings. It provides information on the skills demanded of over 12,000 occupations. The DOT was updated in 1949, 1964, 1977, and 1991, and replaced by the O*NET in 1998.

Given the timing of our study, we use the 1991 DOT, and confirm the results with the 1977 DOT.

The DOT aggregates information into five skill categories. We use these aggregated job task requirements of individual occupations. To link the DOT measures to the CPS and NLSY79 data, we follow Autor, Levy, and Murnane (2003) and use the codes provided on David Autor's website. (Reference: Autor, David H., Frank Levy, and Richard J. Murnane. 2003. "The skill content of recent technological change: An empirical investigation." *Quarterly Journal of Economics* 118 (4): 1279-1333.) We then use the unified IPUMS occupation codes to have consistent coding of occupations of individuals over time. This gives DOT measures for each person-year.

III.B. Industry

To calculate the job task requirements by industry, we use the weighted average job task requirements of workers in the industry. We weight by the number of hours worked (divided by 2000) of each worker.

III.C. Variables

Nonroutine Analytical	The degree to which the task demands analytical flexibility, creativity, reasoning, and generalized problem-solving.
Nonroutine Direction, Control, Planning	The degree to which the task demands complex interpersonal communications such as persuading, selling, and managing others.
Nonroutine Manual	The degree to which the task demands eye, hand, and foot coordination.
Routine Analytical	The degree to which the task requires the precise attainment of set standards,
Routine Manual	The degree to which the task requires repetitive manual tasks