Online Appendix to "MOBILITY AND THE METROPOLIS: THE RELATIONSHIP BETWEEN INEQUALITY IN URBAN COMMUNITIES AND ECONOMIC MOBILITY"

Bryan S. Graham*and
Patrick T. Sharkey†

08 July 2013

1 Overview of NLSY79 Analytic Sample

The National Longitudinal Survey of Youth 1979 (NLSY79) cohort sample contains information on men and women born between 1957 and 1964 inclusive and resident in the United States in 1979 (the first year of data collection). The NLSY79 database is composed of three separate probability samples. The first is a cross sectional sample designed to be representative of the non-institutionalized civilian youth population. The second was designed to oversample Blacks, Hispanics and poor non-Black, non-Hispanic youth. The third was designed to be representative of those aged 17 to 21 at baseline and serving in the military. Funding cutbacks reduced the rate of follow-up of this last group starting in 1985. Frankel and McWilliams (1983) provide a detailed description of the sample design from which we draw on freely below.

A total of 12,686 respondents aged 14 to 22 were successfully interviewed at baseline in 1979. Of those 1,280 were from the military sample. We exclude these units from our analysis both because they represent a specialized subpopulation and because they were not as intensively

*Department of Economics, University of California - Berkeley, 508-1 Evans Hall #3880, Berkeley, CA 94720-3880, e-mail: bgraham@econ.berkeley.edu, web: http : //www.econ.berkeley.edu/ bgraham/

†Department of Sociology, New York University, 295 Lafayette Street, New York, NY 10012, e-mail: patrick.sharkey@nyu.edu, web: http : //sociology.as.nyu.edu/object/patricksharkey.html

followed-up in the later years. This leaves a total 11,406 respondents. The 1,643 poor non-Black, non-Hispanic respondents in the supplementary sample were not interviewed from 1991 onwards (see MaCurdy, Mroz and Gritz, 1998). As, due to truncated follow-up, final educational attainment and adult income is difficult to measure for these respondents we exclude them as well. This leaves a baseline analytic sample consisting of 9,763 respondents.

Of these 9,763 respondents a total 7,272 were aged 19 and under at baseline. For this subgroup we can, at least in principal, measure Metropolitan Statistical Area (MSA) of residence and parental income *during adolescence*. MSAs are defined by the U.S. Office of Management and Budget (OMB). They typically consist of a group of adjacent counties with a common urban core. One way to think about MSAs is as distinct urban labor markets. The definitions of MSAs have changed periodically over the past thirty years. These definitional changes and coding updates makes assembling and merging all relevant MSA data (itself assembled from multiple sources) with the NLSY79 micro data challenging.

Using confidential restricted use geocodes made available to us by special agreement from the Bureau of Labor Statistics (BLS) we matched NLSY79 respondents with Standard Metropolitan Statistical Area (SMSA) of residence at baseline (we used 1981 OMB definitions). A total of 5,120 of the 7,272 respondents aged 19 and under resided in an SMSA at baseline. We refer to these 5,120 respondents as our *core sample*, although at times we will also utilize information from the 2,152 respondents not residing in an SMSA at baseline. What is detailed below is based, unless noted otherwise, on the 5,120 respondents in our core sample.

Education mobility sample

The analysis in the main report uses the "income mobility sample" described below. Here, for completeness, we describe an "education mobility" sample we also constructed in connection with our project. Our income mobility sample is, essentially, a subset of this sample.

At baseline respondents were asked about their parents' education status (or, if the respondents were living with their parents at baseline, the parents' themselves were asked). In the few cases where this information was missing we were sometimes able to successfully extract parental education information from the household roster data available in the 1979, 1980 or 1981 waves of the survey. Using this approach we were able to construct valid measures of highest grade completed for both parents in 4,497 cases, for mothers alone in 523 cases and for fathers alone in 51 cases. For 49 respondents no valid parental education data are available. To summarize we have valid parental education information for at least one parent for 99.0 percent of respondents in our core sample.

We computed measures of parental age in 1979 as follows. First we looked at the parental age questions asked in the 1987 and 1988 rounds of the survey. Second we looked at the parental birthday questions asked during those same rounds. Finally we looked at the age data reported in the household rosters from the 1979, 1980 and 1981 rounds of the survey. We discard any mother age observations that implied an age less than 13 or greater than 50 when the respondent was born. For fathers we discard age measures implying an age of less than 13 or greater than 70 when the respondent was born. Using this approach we get valid age measures for 4,399 respondent fathers and 4,968 respondent mothers (out of 5,120).

For all respondents we attempted to construct a measure of highest grade completed at age 24 (by which time most individuals have completed their formal education). We were able to do this for all but 162 respondents in our core sample (96.8 percent of all cases).

Of the 5,120 respondents in the core sample 3,942, or 77.0 percent, have complete data on birth parents' age and education as well as own education. Summary statistics – means and standard deviations – for these complete case, broken down by race, are given in the following table.

	(1) - All	(2) - Blacks	(3) Hispanics
Years of schooling (YOS) at age 24	12.8 (2.2)	12.7 (1.9)	12.0 (2.3)
Father's schooling in 1979	11.2 (4.0)	10.6 (3.6)	8.6 (4.6)
Father's age in 1979	46.7 (7.5)	46.2 (8.1)	47.0 (8.1)
Mother's schooling in 1979	11.1 (3.2)	11.3 (2.6)	8.3 (4.1)
Mother's age in 1979	43.4 (6.6)	42.6 (6.7)	43.1 (6.8)
Black	0.27	-	-
Hispanic	0.21	-	-
Male	0.50	0.50	0.50
Number of observations	$3,942$	$1,047$	830

Income mobility sample

In order to analyze intergenerational income mobility we require measures of family income during a respondent's childhood. For data availability reasons we focus on measuring family income during a respondent's adolescent years (i.e., between the ages of 13 and 18 inclusive). This is in keeping with most work on intergenerational income mobility (cf, Solon, 1999).

In the NLSY79 parental/family income information is only available for respondents who were co-residing with their parents at the time of interview. Income data collected during each wave corresponds to the prior calendar year. For example the family income data for respondents aged 14 at baseline (in 1979) corresponds to income during the prior year when they were 13. Any unit in core sample should have parent income information for at least one year during their 13 to 18 year old period if they were still residing with their parents at baseline.

Of the 5,120 respondents in our core sample 2,826 still resided with both birth parents, 1,493 with their birth mother, and 199 with their birth father at baseline (totaling 88.2 percent of all respondents). The following table lists the parental co-residence status for core sample respondents between 1979 and 1984. In 1984 income data on 1983 was collected, this corresponds to the year when the youngest members of our sample were 18 years old. At that time only 47.1 percent of respondents were still residing with a parent.

	Both	Mom	Dad	Neither	Roster Inconsistency	Non-Response	Percent with parent(s)
1979	2,826	1,492	199	574	28	0	88.2
1980	2,565	1,350	166	826	8	205	77.7
1981	2,313	1,233	172	1,230	8	164	72.6
1982	2,026	1,099	155	1,628	3	209	64.1
1983	1,751	976	133	2,073	2	185	55.9
1984	1,424	879	109	2,473	2	233	47.1

Were were able to construct measures of family income during adolescence for a total of 4,318 respondents (84.3 percent) in our core sample. In constructing this measure we required the respondent to be living with at least one birth parent and to be between the ages of 13 and 18 during the period in which parental/family income was being measured. We discard any income measurements lower that \$100 (per capita). All income data is deflated to 2010 US dollars using the CPI-U-R (research) index. Our family income measure during adolescence includes an average of 2.5 annual income measures, with minimums and maximums of, respectively 1 and 6.

In order to control for life-cycle effects in our income mobility analysis (e.g., Solon, 1999; Lee and Solon, 2009) we compute a measure of average age of household head over the period for which parental/family income was measured. We define the household head to be the birth father, step-father or mother as available. The average age of household head for the 4,318 respondents with valid family/income data is 45.3 years old.

Those aged 18 in 1978, turned 22 in 1982, the first calendar year in which we are able to measure adult income for some respondents. We observe a child's adult family income annually from 1982 to 1993 and biennially from 1995 to 2007. For respondents aged 19 and under at baseline we have at least one year of adult income data for all but 145 out 5,120 individuals. For over 70 percent of our income mobility sample we observe adult income for 9 or more years (the maximum is 19 years). These income measures are deflated to 2010 dollars and measures less that $100 per capita are discarded.[1]

From the above data we constructed parent-child records of (i) parent income during adolescence, (ii) own income during adulthood (defined as aged 22 and over), (iii) household head's age during adolescence and (iv) own aged for 4,193 of the 5,120 respondents in our target sample (81.9 percent).

Summary statistics, (unweighted) means and standard deviations, for the 4,193 pairs are provided in the table below.

	(1) All	(2) Blacks	(3) Hispanics
Parents/family income during adolescence	59,962 (39,955)	41,556 (30,100)	46,823 (31,073)
Number of parental income measurements	2.5 (1.4)	2.5 (1.3)	2.5 (1.3)
Average age of household head during adolescence	45.3 (7.4)	44.1 (7.8)	45.4 (7.8)
Number of own income measurements	11.7 (4.3)	10.7 (4.3)	11.3 (4.2)
Average child age during adulthood	31.3 (2.6)	31.5 (2.9)	31.3 (2.6)
Black	0.31	-	-
Hispanic	0.22	-	-
Male	0.52	0.52	0.53
Number of observations	4,193	1,294	930

2 Overview of NLSY97 sample

The National Longitudinal Survey of Youth 1997 (NLSY97) cohort sample contains information on men and women born between 1980 and 1984 inclusive and resident in the United States in 1997 (the first year of data collection). NLSY97 respondents have been interviewed annually since baseline. The latest year for which data are publicly available is 2009 (during which earnings in 2008 are recorded). Note that in 2009 respondents were between 24 and

[1] Note we use measures of family income in our mobility analysis as opposed to, say, father's wage earnings. See Mazumder (2005) and Lee and Solon (2009) for some discussion of the pros- and cons- of this approach.

29 years of age. While the NLSY79 cohort is old enough that we are able to observe earnings realizations at many points in the life cycle (i.e., as late as a respondent's early 50s), the NLSY97 cohort is not yet old enough to make such extensive follow-up possible.

The NLSY97 database is composed of two separate probability samples. The first is a cross sectional sample designed to be representative of the non-institutionalized civilian youth population. The second was designed to oversample Blacks and Hispanics youth. The overall structure of the NLSY97 closely mirrors that of the NLSY79. This close correspondence between the two surveys will allow us to explore how MSA-specific measures of intergenerational mobility vary across two different birth cohorts. The NLSY97 User's Guide describes the database in detail (U.S. Department of Labor, 2003).

A total of 8,985 youth were interviewed at baseline. Using confidential restricted use geocodes made available to us by special agreement from the Bureau of Labor Statistics we matched NLSY99 respondents with Metropolitan Statistical Area (MSA) of residence at baseline (we used 1999 MSA definitions).[2] A total of 7,263 of the 8,985 respondents resided in an MSA at baseline. We refer to these 7,263 respondents as our *core sample*, although at times we will also utilize information from the 1,722 respondents not residing in an MSA at baseline. What is detailed below is based, unless noted otherwise, on the 7,263 respondents in our core sample.

Education mobility sample

As with our NLSY79 analysis, the NLSY97 analysis in the main report uses the "income mobility sample" described below. Here, for completeness, we describe an "education mobility" sample we also constructed in connection with our project. Our income mobility sample is, essentially, a subset of this sample.

[2] For New England residents we used New England Consolidated Metropolitan Areas (NECMAs).

Respondents (or their parents if applicable) were asked about the educational attainment of their biological parents. Parental education data is also recorded on the household residential and non-residential rosters at baseline, and in the 1998 residential roster. Using these various survey items we were able to successfully measure father's years of completed schooling for 81.0 percent of respondents. Mother's years of completed schooling was successfully measured for 92.8 percent of respondents. For 78.6 percent of respondents ($N = 5,708$) we observed years of completed schooling for both biological parents. For 95.2 percent of respondents we observed years of completed schooling for at least one parent.

We measure parents' age at baseline using the date of birth question in the 1997 household roster. If this item is missing we use the information in the age question (which results in a coarser measure – age in years versus months). Third we extract information from the 1997 non-residential roster. For a small number of parents we extract age at baseline from information contained in the 1998 and 1999 residential and non-residential rosters. We discard all responses which imply an age less that 13 and greater than 50 for mothers at the time of the respondent's birth. For fathers we discard units will ages below 13 and in excess of 70 at respondent's birth. Using this approach we generated valid measures of father's and mother's age for, respectively, 87.5 and 96.8 percent of respondents.

We measure respondent's years of completed schooling as of age twenty four. We were able to successfully compute this measure for 89.3 percent of respondents, with missing outcomes due primarily to attrition from the sample (i.e., not item non-response).

Of the 7,263 respondents in the core sample 4,945, or 68.1 percent, have complete data on birth parents' age and education as well as own education. Summary statistics for these complete case, broken down by race, are given in the following table.

	(1) - All	(2) - Blacks	(3) Hispanics
Years of schooling (YOS) at age 24	13.4 (2.6)	13.0 (2.5)	12.8 (2.4)
Father's schooling in 1997	12.5 (3.5)	12.3 (2.4)	10.0 (4.2)
Father's age in 1997	43.1 (6.6)	41.9 (7.3)	42.9 (6.9)
Mother's schooling in 1997	12.6 (3.1)	12.7 (2.0)	10.3 (3.8)
Mother's age in 1997	40.7 (5.4)	39.5 (5.7)	40.3 (5.6)
Black	0.22	-	-
Hispanic	0.24	-	-
Male	0.51	0.48	0.52
Number of observations	$4,945$	$1,104$	$1,198$

Income mobility sample

In order to analyze intergenerational income mobility we require measures of family income during a respondent's childhood. For data availability reasons we focus on measuring family income during a respondent's adolescent years (i.e., between the ages of 11 and 18 inclusive). This is in keeping with most work on intergenerational income mobility (cf, Solon, 1999). Relative to our analysis with the NLSY79 cohort we are able to include family income measures for earlier years for some respondents since due to differences in the age composition of respondents at baseline across the two surveys.

As in the NLSY79 parental/family income information is available for respondents who were co-residing with their parents at the time of interview. Income data collected during each wave corresponds to the prior calendar year. For example the family income data for respondents aged 12 at baseline (in 1997) corresponds to income during the prior year when they were 11.

Of the 7,263 respondents in our core sample, a total of 3,532 still resided with both birth parents, 2,892 with their birth mother, and 409 with their birth father at baseline. A total of 94.1 percent of respondents resided with at least one biological parent at baseline.

The following table lists the parental co-residence status for core sample respondents between 1997 and 2004. In 2004 income data for 2003 was collected, this corresponds to the year when the youngest members of our sample were 18 years old. At that time only 37.6 percent of respondents still resided with a birth parent.

	Both	Mom	Dad	Neither	Roster Inconsistency	Non-Response	Percent with parent(s)
1997	3,532	2,892	409	430	0	0	94.1
1998	3,186	2,539	409	582	7	540	84.5
1999	3,028	2,349	391	825	7	663	79.4
2000	2,726	2,177	361	1,207	4	788	72.5
2001	2,500	1,942	329	1,538	4	950	65.7
2002	2,185	1,745	289	2,119	2	923	58.1
2003	1,743	1,395	243	2,839	8	1,035	46.6
2004	1,387	1,142	203	3,309	4	1,218	37.6

Were were able to construct measures of family income during adolescence for a total of 6,239 respondents (85.9 percent of the core sample). In constructing this measure we required the respondent to be living with at least one birth parent and to be between the ages of 11 and 18 during the period in which parental/family income was being measured. We discard any income measurements lower that $100 (per capita). All income data is deflated to 2010 US dollars using the CPI-U-R 'research' index. Our family income measure during adolescence includes an average of 2.1 annual income measures, with minimums and maximums of, respectively 1 and 7.

In order to control for 'life cycle' bias (e.g., Grawe, 2006) we constructed a measure of the (average) age of the household head over the years in which parental/family income was being averaged over. We define the household head to be the father, step-father or mother as appropriate. The parental age measures were constructed as described above. The age of the step-father, if present, was constructed using information contained in each wave's household roster.

We observe the family income of NLSY97 respondents in later waves as they age into adulthood. We make these measurements from age 22 onwards. We were able to construct at least one valid adult income measure for 81.3 percent of respondents and as many as seven for some older respondents. As before we discard observations implying per capita income less that $100 and deflate all income measures to 2010 dollars.

The following table summarizes some basic features of the sub-sample of 5,537 parent-child pairs for which were able to measure family income during both adolescence (while residing with at least one biological parent) and during adulthood. This subsample corresponds to 76.2 percent of our core sample. Unweighted means and standard deviations are reported.

	(1) - All	(2) - Blacks	(3) Hispanics
Parents/family income during adolescence	77,864 (65,897)	53,387 (47,654)	56,470 (47,753)
Number of parental income measurements	2.2 (1.0)	2.1 (1.0)	2.1 (1.0)
Average age of household head during adolescence	43.5 (6.8)	41.8 (7.0)	43.4 (6.9)
Number of own income measurements	3.4 (1.5)	3.3 (1.5)	3.4 (1.6)
Average age during adulthood	24.2 (0.9)	24.2 (1.0)	24.2 (0.9)
Black	0.25	-	-
Hispanic	0.25	-	-
Male	0.51	0.48	0.51
Number of observations	5,537	1,410	1,372

3 Overview of the PSID Sample

The PSID is an ongoing longitudinal survey begun in 1968 with a total sample of about 4,800 families (Hill and Morgan, 1992). Interviews of sample families were conducted on an annual basis from 1968 through 1997, when the PSID switched to a biennial survey. Our file includes all survey years from 1968 through 2007. The PSID follows all family members of the original sample as they "splitoff" from the sample family, whether through some family disruption such as divorce or when children enter adulthood and form their own families. This feature of the survey allows for analysis of intergenerational relationships.

The PSID is composed of multiple samples. First, the "Survey Research Center" (SRC) sample is a nationally representative sample of American households as of 1967. Second, the "Survey of Economic Opportunity" (SEO) sample is an oversample of low-income households (Hill and Morgan, 1992). A document describing the selection of the SEO sample notes several uncertainties about the procedures that were used to select the original sample (Brown, 1996). However, several subsequent studies have shown that the SEO sample is representative of the low-income population when compared with the Current Population Survey (Becketti et al., 1988; Fitzgerald, Gottschalk, and Moffitt, 1998a; Fitzgerald, Gottschalk, and Moffitt, 1998b). Our analysis files use both the SRC and SEO samples. Several decades after the PSID began additional samples were included in order to make the survey representative of the changing population in the U.S. We exclude these samples because only a small number of such sample members have been in the survey for a sufficient number of years to appear in the data as children and again as adults.

The multigenerational data file from the PSID begins with a sample of children from PSID sample families and follows these children into adulthood. Whereas the NLSY samples measure childhood characteristics only in adolescence, the extended duration of the PSID allows us to measure childhood characteristic s over all years in which the individual is classified as a child in a PSID sample family, biological or adoptive, from age 0 through 17. We measure all adult characteristics over all years in which the individual is classified as the household head or the spouse of a household head and is at least 26 years old.

There are 27,383 individuals who are ever identified as "children" in a PSID sample household from the age of 0 through 17. Of this group, there are 7,336 individuals who are also observed as household heads or the spouses of household heads in at least one year at the age of 26 or older. The oldest sample members who are observed as children and adults were born in 1951, and were observed as 17 year olds in 1968, the first year of the survey. The youngest sample members who are observed as children and adults were born in 1981, and were age 26 in the 2007 survey. Considering this is a study of economic mobility across metropolitan areas, we limit the sample to include individuals who lived within a metropolitan area in at least one survey year during childhood. We also limit the sample to metropolitan areas with at least 100,000 residents as of the 2000 Census, and we measure mobility only in MSAs that have at least 25 PSID households. This reduces the sample to 4,320 individuals.

About 90 percent of children in the PSID sample live in only one metropolitan area during childhood. For those who live in more than one metropolitan area at some point during childhood, we define the individual's metropolitan area as the one in which the individual spent the most years during childhood. If there is a tie, we use the last metropolitan area in which the individual is observed during childhood. Although we use the general term "metropolitan area" in this memo, urban areas are operationalized using Primary Metropolitan Statistical Areas (PMSAs) and Metropolitan Statistical Areas (MSAs) as defined by the Census Bureau. PMSA/MSAs are areas with a core urban area with 50,000 or more inhabitants along with adjacent communities that have a "high degree" of economic and social integration with the core urban area.

Family income is measured as the total amount of taxable income plus transfer income among the household head, the spouse of the household head, and all other members of the family unit in the year prior to the survey. The measure of family income is inflated to represent 2010 dollars using the Consumer Price Index – Research Series (CPI-U-RS). To maintain consistency with the coding of survey responses in the early years of the survey, we have recoded all income values less than or equal to \$0 to \$1. Family income is measured during childhood and again in adulthood. The measure of childhood family income is an average over all years in which the individual is observed as a child in a PSID household between the ages of 0 and 17, inclusive. We discard cases where average family income is less than \$100. The measure of adult family income is an average over all years in which the individual is a household head or spouse of a household head and is at least 26 years old.

4 Construction of Neighborhood Sorting Index (NSI)

The Neighborhood Change Database (NCDB) provides information, at the census tract level, on the number of households in each of a set of mutually exclusive income bins. The number and size of these bins varies across census years. In 1990, for example, there are a total of 19 bins, ranging from \$0 to \$4,999 at the bottom to greater than or equal to \$150,000 at the top. Let $T+1$ denote the number of available income bins. Let inc_t denote the upper end point of the t^{th} bin. Let i index census tracts, and h households. Let $p_{i,t} = \Pr\left(\text{INC}_{ih} \leq \text{inc}_t \middle| i\right)$ be the probability that a randomly sampled household *in census tract* i has an income less than or equal to inc_t as reported in the NCDB.

Our basic approach to computing the NSI extends that used by Wheeler and La Jeunesse (2008). We decompose the logarithm of household income into a between- and within-neighborhood component:

$$y_{ih} = \ln\left(\text{INC}_{ih}\right) = \alpha_i + \varepsilon_{ih},$$

where α_i is mean log-income in neighborhood i and ε_{ih} household h's deviation from that mean. Assuming ε_{ih} is normally distributed, or, equivalently, that the within-neighborhood distribution of household income is lognormal, we have

$$\text{inc}_\text{t} = \exp\left(\alpha_i + \sigma_\varepsilon \Phi^{-1}\left(p_{i,t}\right)\right),$$

for $t = 1, \ldots, T$ and $\Phi^{-1}\left(\bullet\right)$ the inverse CDF of a standard normal random variable.

Let $i = 1, \ldots, N$, so that N equals the total number of census tracts in the MSA. We compute σ_ε as the coefficient on $\Phi^{-1}\left(p_{i,t}\right)$ in the least squares fit of $\ln\left(\text{inc}_t\right)$ onto a vector of tract-specific dummy variables and $\Phi^{-1}\left(p_{i,t}\right)$. This fit is computed using the T quantile values given for each of the N tracts (i.e., using a total of NT 'observations'). This approach to construction an estimate of σ_ϵ appears to be new. The R^2 associated with these OLS fits is typically quite higher, averaging above 0.95.

To compute the variance of α_i we aggregate the NCDB tract counts up to the MSA-level. We then compute the least squares fit of $\ln\left(\text{inc}_t\right)$ onto a constant and $\Phi^{-1}\left(p_t\right)$ where p_t is the proportion of households *across the entire MSA* with an income less than or equal to inc_t. This MSA-specific fit is computed using T 'observations'. The coefficient on $\Phi^{-1}(p_t)$ is our estimate of $\sqrt{\sigma_\alpha^2 + \sigma_\varepsilon^2}$. Wheeler and La Jeunesse (2008) previously used this approach to construct an estimate of the standard deviation of log income at the MSA level. The R^2 associated with these OLS fits is also very high.

Our estimate of the neighborhood sorting index (NSI) is one minus the square of the ratio of the first to the second coefficient estimates. This procedure assumes log normality of the income distribution at the MSA-level. The high R^2 values associated with our two sets of OLS fits, suggests that our log-normal assumption is able to replicate the income shares reported in the NCDB quite accurately, although their is some concern that the fit may be poorer at higher percentiles of the income distribution (Jargowsky, 1996).

5 Additional details on geocoding and MSA definitions

U.S. Office of Management and Budget MSA definitions have varied over time. Since one goal of our analysis is to conduct analysis of how measured mobility within a given metropolitan area has changed over time these changing MSA boundaries are potentially problematic. A further complication is caused by the differing treatment of MSAs within New England. For these reasons we mapped the 1981 SMSA and 1999 MSA/NECMA definitions into a common set of self-defined 259 Merged Metropolitan Statistical Areas (MMSAs). Typically this involved aggregating several 1981 SMSA into a larger 1999 units or vice-versa. For example in 1981 the OMB treated "Denver-Boulder" as a single metropolitan area, while in 1999 these two areas were considered separate MSAs. In this case we assigned all respondents in the NLSY79 living in either Denver or Boulder at baseline to the common Denver-Boulder MMSA. A spreadsheet summarizing our concordance is available online (the filename is `MSA81To99Concordance.csv`). In most cases our MMSA definitions mirror OMBs 1999 ones. All of our analysis uses our MMSA definitions.

6 Empirical framework

The analysis reported in our main report is straightforward. City-specific estimates of the intergenerational elasticity of family income, a widely-used measure of mobility, are correlated with various MSA-level variables using weighted least squares techniques. The weights are proportional to the inverse of the sampling error of the city-level IGE estimates (i.e., more precisely estimated IGEs are given greater weight). This section develops some additional, more technical, properties of our empirical framework that may be of interest to some readers.

Individuals "grow up" in different cities indexed by c. Within cities they, or rather their parents, array themselves into different neighborhoods indexed by n. We write $Y_{t,inc}$ for the i^{th} child's outcome from neighborhood n in city c (e.g., highest grade completed or the logarithm of net family income). We use $Y_{t-1,inc}$ to denote the corresponding outcome for the child's parent(s). Define $\bar{Y}_{t-1,nc} = H_{nc}^{-1} \sum_{i=1}^{H_{nc}} Y_{t-1,inc}$ to be the neighborhood average of the parental outcome, where H_{nc} is the number of households in the n^{th} neighborhood of city c. Additional unmeasured child characteristics are captured by the latent variable $A_{t,inc}$ which we refer to as 'ability' in a shorthand.

Let the best linear predictor[3], *conditional on residence in city c,* of child's outcome given parents' outcome, neighborhood average parental outcome and ability be

$$\begin{aligned}\mathbb{E}^* \left[Y_{t,inc} \middle| Y_{t-1,inc}, \overline{Y}_{t-1,nc}, A_{t,inc}; c \right] &= \alpha(c) + \beta(c) Y_{t-1,inc} \\ &\quad + \gamma(c) \bar{Y}_{t-1,nc} + \delta(c) A_{t,inc}. \end{aligned} \tag{1}$$

Equation (1) allows the (average) mapping from parental background, neighborhood 'quality' and 'ability' into own outcomes to vary across cities. The notation $\alpha(c)$, for example, denotes that the intercept in (1) may vary across cities.[4] It may be that certain cities are more or less mobile for any number of reasons. Possibilities include differences in educational systems, demographic and/or economic structure.

Because (1) conditions on the latent variable $A_{t,inc}$ we view it as a structural or causal relationship. Specifically $\beta(c)$ and $\gamma(c)$ measure the causal effect of a unit increase in parental and neighborhood 'quality' *in city* c on own adult outcomes (cf., Wooldridge, 2005).

In practice estimating the parameters of (1) directly is infeasible since ability is unobserved. Instead we can compute the best linear predictor of own outcome on parents' outcome alone (again conditioning on city of residence). Using (1) we have

$$\begin{aligned}\mathbb{E}^* \left[Y_{t,inc} \middle| Y_{t-1,inc}; c \right] &= \alpha(c) + \beta(c) Y_{t-1,inc} \\ &\quad + \gamma(c) \mathbb{E}^* \left[\bar{Y}_{t-1,nc} \middle| Y_{t-1,inc}; c \right] \\ &\quad + \delta(c) \mathbb{E}^* \left[A_{t,inc} \middle| Y_{t-1,inc}; c \right]. \end{aligned} \tag{2}$$

To derive an interpretable expression for (2) we need to develop some notation for the two

[3] A mean square error minimizing linear predictor is simply the population analog of a least squares fit. An alternative terminology would be the linear regression function. Goldberger (1991) provides a textbook introduction to LPs and Chamberlain (1984) applications of them to the analysis of panel data.

[4] Formally (1) is a conditional linear predictor (CLP). See Wooldridge (1999) for some properties of CLPs.

linear predictors to the right of the equality. Denote the best linear predictor of own ability given parents' background in city c by

$$\mathbb{E}^{*}\left[A_{t,inc}\right|Y_{t-1,inc};c]=\zeta\left(c\right)+\eta\left(c\right)Y_{t-1,inc}. \tag{3}$$

Our expression for the best linear predictor of neighborhood quality given parents' background exploits an algebraic relationship connecting the within- and between-neighborhood variation in parents' background. Specifically we can show after some tedious manipulation that (see Section 7 below)

$$\mathbb{E}^{*}\left[\bar{Y}_{t-1,nc}\middle|Y_{t-1,inc};c\right]=\left\{\mu_{\overline{Y}}\left(c\right)-\mu_{Y}\left(c\right)\mathrm{NSI}\left(c\right)\right\}+\mathrm{NSI}\left(c\right)\times Y_{t-1,inc}, \tag{4}$$

where $\mu_{\overline{Y}}\left(c\right)=\mathbb{E}\left[\bar{Y}_{t-1,nc}\middle|c\right]$, $\mu_{Y}\left(c\right)=\mathbb{E}\left[Y_{t-1,inc}\right|c]$, and $\mathrm{NSI}\left(c\right)$ is the neighborhood sorting index (NSI) of Farley (1977, p. 503) and others

$$\mathrm{NSI}\left(c\right)=\frac{\mathbb{V}\left(\sqrt{\frac{H_{nc}}{\mu_{H}(c)}}\times\bar{Y}_{t-1,nc}\middle|c\right)}{\mathbb{V}\left(Y_{t-1,inc}\right|c)}, \tag{5}$$

or the ratio of the between-neighborhood variance in parental background to the overall variance (in city c); here $\mu_{H}\left(c\right)=\mathbb{E}\left[H_{nc}\right|c]$ denotes the average neighborhood size in city c.[5] The weight in the numerator of (5) adjusts for heterogeneity in neighborhood size, up-weighting larger neighborhoods and down-weighting smaller ones. The sample counterpart of (5) equals the ratio a standard one-way ANOVA estimate of the between-neighborhood variance in $Y_{t-1,inc}$ to its overall variance (cf., Theorem 11.2.4 of Casella and Berger (1990, p. 525)).

Under random assignment of families to neighborhoods (5) will equal zero. In the other extreme, where neighborhoods are perfectly stratified by parental background, it will equal 1. These statements assume the existence of a continuum of neighborhoods and are only approximate when, as is the case in fact, there are only a finite number of neighborhoods.

Plugging equations (3) and (4) into (2) gives

$$\mathbb{E}^{*}\left[Y_{t,inc}\right|Y_{t-1,inc};c]=a\left(c\right)+b\left(c\right)Y_{t-1,inc} \tag{6}$$

[5] Different authors define the NSI in different ways. For example Jargowsky (1996) calls the square root of (5) the neighborhood sorting index (p. 988). We use the above definition throughout.

where

$$\begin{aligned}
a\left(c\right) &= \alpha\left(c\right)+\left\{\mu_{\overline{Y}}\left(c\right)-\mu_{Y}\left(c\right)\mathrm{NSI}\left(c\right)\right\}+\zeta\left(c\right)\\
b\left(c\right) &= \beta\left(c\right)+\gamma\left(c\right)\times\mathrm{NSI}\left(c\right)+\delta\left(c\right)\times\eta\left(c\right).
\end{aligned}$$

Equation (6) indicates that the intergenerational elasticity (IGE) in earnings (or family income, education etc.) in city c, here denoted by $b\left(c\right)$, may be decomposed into three components. The first component $\beta\left(c\right)$ is the causal effect of parents' background on own outcomes. The second component, $\gamma\left(c\right)\times\mathrm{NSI}\left(c\right)$, captures the effect of neighborhood influence and stratification on mobility. The third component, $\delta\left(c\right)\times\eta\left(c\right)$, captures the influence of any correlation between unmeasured attributes and parental background on measured mobility.

The last component of the IGE has been widely discussed in the sociology and economics literature on mobility (see Solon (1999) for a review). Its presence accounts for the fact that most authors view the IGE as a descriptive as opposed to a causal parameter (e.g., Mazumder, 2005).

The second component of the slope coefficient in (6) was intuitively described in Loury (1977) and is also an implication of the formal theoretical models developed by Durlauf (1996) and Graham (2008). The intuition underlying this component of the IGE is straightforward to explain. Imagine a city with no stratification by parental background across neighborhoods. In such a city parental background will be uncorrelated with neighborhood quality. Since parental background has no predictive power for neighborhood quality in such a city the IGE will capture the private effect of background (and any ability bias) alone. Now consider a city with a high level of stratification by parental background across neighborhoods. In such a city parents' background will be highly predictive of neighborhood quality. In that case the IGE will capture both the private effect of parental background as well as the social effect operating via neighborhood influences, $\gamma\left(c\right)$, and sorting $\mathrm{NSI}\left(c\right)$. In such a city having a parent from with low educational attainment, for example, suggests that one also grew up in a neighborhood with other adults with low education levels. In the limit, where neighborhoods are perfectly stratified by parental background we have $\mathrm{NSI}\left(c\right)=1$ so that

$$b\left(c\right)=\beta\left(c\right)+\gamma\left(c\right)+\delta\left(c\right)\times\eta\left(c\right).$$

Decomposition (6) has implications for the observed variation in intergenerational mobility across cities. Specifically, *holding all other city characteristics constant,* we should expect the intercept of a simple bivariate regression of child's outcome on parents' outcome to be decreasing in the degree of residential stratification in the city (i.e., in $\mathrm{NSI}\,(c)$) and the slope coefficient to be increasing in this quantity (assuming, as seems reasonable, that $\gamma\,(c) > 0$). Unfortunately, at least from the perspective of our research design, cities are heterogeneous in many ways and these differences may exert their own effect on measured mobility independent on any effect operating through neighborhood quality. As one of many examples the market price of ability, here captured by the coefficient on ability in (1), $\delta\,(c)$, could differ across cities. This price variation may in turn be correlated with the observed degree of residential stratification (e.g., the rich sort into certain high price neighborhoods for their amenity value alone). In such a situation we might observe a correlation between measured mobility and stratification across cities (i.e, between $b\,(c)$ and $\mathrm{NSI}\,(c)$) even in the absence of any human capital benefits of neighborhood quality.

We have identified three approaches to addressing biases caused by unmeasured heterogeneity across cities. The first and perhaps simplest approach is to condition the analysis on additional city-level characteristics. Let X_c be a vector of such characteristics. For simplicity, assume that $\gamma\,(c)$, $\delta\,(c)$ and $\eta\,(c)$ are constant across cities and equal to, respectively, γ_0, δ_0 and η_0. The research designs outlined below can be applied without this assumption, but making it serves to sharpen our discussion and simplify the expressions that follow. We write, for cities $c = 1, \ldots, C$

$$\mathbb{E}^*\left[\left. b\,(c)\right| X_c, \mathrm{NSI}\,(c)\right] = \beta_0 + \delta_0\eta_0 + \gamma_0 \mathrm{NSI}\,(c) + \mathbb{E}^*\left[\left.\beta\,(c) - \beta_0\right| X_c, \mathrm{NSI}\,(c)\right],$$

where $\beta_0 = \mathrm{E}\left[\beta\,(c)\right]$. If

$$\mathbb{E}^*\left[\left.\beta\,(c) - \beta_0\right| X_c, \mathrm{NSI}\,(c)\right] = \mathbb{E}^*\left[\left.\beta\,(c) - \beta_0\right| X_c\right] = \phi_0 + X_c'\pi_0 \tag{7}$$

so that measured stratification is uncorrelated with cross city variation in the effect of parental background on child outcomes *conditional* on X_c then the coefficient on $\mathrm{NSI}\,(c)$ in a least squares fit of $b\,(c)$ onto a constant, $\mathrm{NSI}\,(c)$ and X_c will be consistent for γ_0. Assumption (7) is very strong, but this approach is straightforward to execute and understand: does measured stratification predict measured mobility conditional on a vector basic city characteristics?

A second approach exploits variation in measured mobility across two cohorts raised in the same city but during different time periods. Let $b_b\,(c)$ denote measured mobility in city c for birth cohort $b = 1, 2$ and decompose $\beta_b\,(c) = \bar{\beta}\,(c) + \rho_b + \varepsilon_b\,(c)$ into a city-specific, cohort-specific and city-by-cohort components. Denoting $\triangle b\,(c) = b_2\,(c) - b_1\,(c)$ we have

$$\mathbb{E}^*\left[\left.\triangle b\,(c)\right| X_c, \triangle\mathrm{NSI}\,(c)\right] = \rho_2 - \rho_1 + \gamma_0 \triangle\mathrm{NSI}\,(c) + \mathbb{E}^*\left[\left.\triangle\varepsilon_b\,(c)\right| X_c, \mathrm{NSI}\,(c)\right].$$

If

$$\mathbb{E}^{*}\left[\left.\triangle\varepsilon_{b}\left(c\right)\right|X_{c},\mathrm{NSI}\left(c\right)\right]=\mathbb{E}^{*}\left[\left.\triangle\varepsilon_{b}\left(c\right)\right|X_{c}\right]=\phi_{0}+X_{c}^{\prime}\pi_{0},\tag{8}$$

then the coefficient on $\triangle\mathrm{NSI}\left(c\right)$ in a least squares fit of $\triangle b\left(c\right)$ onto a constant, $\triangle\mathrm{NSI}\left(c\right)$ and X_{c} will consistently estimate γ_{0}. Condition (8) requires that within-city *changes* in measured residential stratification across cohorts are conditionally uncorrelated with other changes in other city-specific determinants of intergenerational mobility. Note that the vector X_{c} need not coincide in (7) and (8).

A final approach posits the existence of instrumental variables. Specifically we write

$$b\left(c\right)=\beta_{0}+\delta_{0}\eta_{0}+\gamma_{0}\mathrm{NSI}\left(c\right)+\left\{\beta\left(c\right)-\beta_{0}\right\},$$

then if we can find a variable Z_{c} that is correlated with $\mathrm{NSI}\left(c\right)$ but uncorrelated with $\left\{\beta\left(c\right)-\beta_{0}\right\}$ we may apply the method of two-stage least squares (TSLS) to consistently estimate γ_{0}. Plausible instruments for $\mathrm{NSI}\left(c\right)$ include geographic features of a city as well as policy variables which influence the distribution of human capital in the parents' generation.

In the main report we develop the first two approaches empirically, although, in our view, neither assumptions (7) or (8) are completely plausible and our analysis is consequently correlational in nature. Nevertheless the above framework provides some insight into what a structural interpretation of our results would entail.

7 Derivations

Equation (4) follows from the formula for the best linear predictor and the orthogonality of the within- and between-neighborhood variation in parental background. A subtlety arises because the coefficient on $Y_{t-1,inc}$ in (4) is a function of moments computed with respect to the population of *individuals* within city c, while the neighborhood sorting index involves moments computed with respect to the population of *neighborhoods*.

Specifically the slope coefficient in (4) is derived as follows:

$$\begin{aligned}
\frac{\mathbb{C}\left(Y_{t-1,inc}, \bar{Y}_{t-1,nc}\right| c\right)}{\mathbb{V}\left(Y_{t-1,inc}\right| c\right)} &= \frac{\mathbb{C}\left(\bar{Y}_{t-1,nc} + \left(Y_{t-1,inc} - \bar{Y}_{t-1,nc}\right), \bar{Y}_{t-1,nc}\right| c\right)}{\mathbb{V}\left(Y_{t-1,inc}\right| c\right)} \\
&= \frac{\mathbb{C}\left(\frac{H_{nc}}{\mu_H(c)}\bar{Y}_{t-1,nc}, \bar{Y}_{t-1,nc}\right| c\right) + \mathbb{C}\left(\left(Y_{t-1,inc} - \bar{Y}_{t-1,nc}\right), \bar{Y}_{t-1,nc}\right| c\right)}{\mathbb{V}\left(Y_{t-1,inc}\right| c\right)} \\
&= \frac{\mathbb{V}\left(\sqrt{\frac{H_{nc}}{\mu_H(c)}}\bar{Y}_{t-1,nc}\right| c\right) + 0}{\mathbb{V}\left(Y_{t-1,inc}\right| c\right)} \\
&= \mathrm{NSI}\left(c\right).
\end{aligned}$$

Note that in the first line of the above derivation all moments are defined with-respect to the population of individuals. The first term in the numerator of the second line, in contrast, is defined with respect to the population of neighborhoods. The introduction of the weighting factor in this expression is to ensure numerical equivalency with expression on the right-hand-side of the first line. The third equality follows from orthogonality of the within- and between-neighborhood variation in parental background and the definition of variance.

References

[1] Becketti, Sean, William Gould, Lee Lillard, and Finis Welch. (1988). "The Panel Study of Income Dynamics after Fourteen Years: An Evaluation." *Journal of Labor Economics* 6 (4): 472 - 492.

[2] Casella, George and Roger L. Berger. (1990). *Statistical Inference.* Belmont, CA: Duxbury Press.

[3] Chamberlain, Gary. (1984). “Panel data,” *Handbook of Econometrics* 2: 1247-1318 (Z. Griliches & M. Intriligator, Eds.). Amsterdam: North Holland.

[4] Durlauf, Steven N. (1996). “A theory of persistent income inequality,” *Journal of Economic Growth* 1 (1): 75 - 93.

[5] Farley, Reynolds. (1977). “Residential segregation in urbanized areas of the United States in 1970: an analysis of social class and racial differences,” *Demography* 14 (4): 497 - 518.

[6] Fitzgerald, John, Peter Gottschalk, and Robert Moffitt. (1998a). "An Analysis of Sample Attrition in Panel Data: The Michigan Panel Study of Income Dynamics," *Journal of Human Resources* 33 (2): 251 - 299.

[7] Fitzgerald, John, Peter Gottschalk, and Robert Moffitt. (1998b). "An Analysis of the Impact of Sample Attrition on the Second Generation of Respondents in the Michigan Panel Study of Income Dynamics." *Journal of Human Resources* 33 (2): 300 - 344.

[8] Frankel, Martin R. and Bruce D. Spencer. (1983). *National Longitudinal Survey of Labor Force Behavior Youth Survey (NLS): Technical Sampling Report.* Columbus, OH: NLS User Services.

[9] Goldberger, Arthur S. (1991). *A Course in Econometrics.* Cambridge, MA: Harvard University Press.

[10] Graham, Bryan S. (2008). "Endogenous neighborhood selection, the distribution of income, and the identification of neighborhood effects," *Mimeo*, University of California - Berkeley.

[11] Grawe, Nathan. (2006). "Lifecycle bias in estimates of intergenerational earnings persistence," *Labour Economics* 13 (5): 551 - 570.

[12] Hill, Martha S. and James N. Morgan. (1992). *The Panel Study of Income Dynamics: A User's Guide*. Newbury Park, CA: Sage Publications.

[13] Jargowsky, Paul A. (1996). "Take the money and run: economic segregation in U.S. Metropolitan areas," *American Sociological Review* 61 (6): 984 - 998.

[14] Lee, Chul-In and Gary Solon. (2009). "Trends in intergenerational mobility," *Review of Economics and Statistics* 91 (4): 766 - 772.

[15] Loury, Glenn C. (1977). "A dynamic theory of racial income differences," *Women, Minorities and Employment Discrimination*: 153 - 186 (P.A. Wallace & A LeMond, Eds.). Lexington, MA: Lexington Books.

[16] MaCurdy, Thomas, Thomas Mroz, and R. Mark Gritz. (1998). "An evaluation of the National Longitudinal Survey of Youth," *Journal of Human Resources* 33 (2): 345 - 436.

[17] Mazumder, Bhashkar. (2005). "Fortunate sons: new estimates of intergenerational mobility in the United States using social security earnings data," *Review of Economics and Statistics* 87 (2): 235 - 255.

[18] Solon, Gary. (1999). "Intergenerational mobility in the labor market," *Handbook of Labor Economics* 3A: 1761 - 1800 (O. C. Ashenfelter & D Card, Eds.). Amsterdam: North-Holland.

[19] U.S. Department of Labor. (2003). *NLSY97 Users Guide.* Washington D.C.: Bureau of Labor Statistics, U.S. Department of Labor.

[20] Wheeler, Christopher H. and Elizabeth A. La Jeunesse. (2008). "Trends in neighborhood income inequality in the U.S.: 1980-2000," *Journal of Regional Science* 48 (5): 879 - 891.

[21] Wooldridge, Jeffrey M. (1999). "Distribution-free estimation of some nonlinear panel data models," *Journal of Econometrics* 90 (1): 77 – 97.

[22] Wooldridge, Jeffrey M. (2005). "Unobserved heterogeneity and estimation of average partial effects," *Identification and Inference for Econometric Models: Essays in Honor of Thomas Rothenberg*: 27 - 55 (D.W.K. Andrews & J.H. Stock, Eds.). Cambridge: Cambridge University Press.