Cracks in the Melting Pot: Immigration, School Choice, and Segregation

Elizabeth U. Cascio and Ethan G. Lewis

Online Appendix A: Data

A. School District Level Data: Sources and Construction of Key Variables

1. 1970 and 2000 School District Tabulations

For 1970, school district level data on total population, by age and ethnicity, and private school enrollment, by ethnicity and level, were drawn from the 1970 Fourth Count (Population) School District Data Tapes (National Center for Education Statistics (NCES) 1970).[1] These data permit identification of all school districts in the country with at least 300 students as of the 1969-70 school year. For 2000, school district level data on total population and private school enrollment, by age and ethnicity, were drawn from the Census 2000 School District Tabulation (NCES 2000).[2] All operating districts are included in the age-specific resident counts, but private enrollment counts are missing for districts with 49 or fewer children.

Presentation of the data differs across years. For consistency over time in the definition of our key dependent variable, we aggregate non-Hispanic resident counts to the 0-19 and 20-49 age groups and aggregate these counts to constant secondary district boundaries.[3] To arrive at the dependent variable used in our analysis, we divide by the non-Hispanic MSA population for that age group, generated from county level Census data available at the National Historical Geographic Information System (Minnesota Population Center 2004). We use a similar approach to calculating the non-Hispanic private school enrollment rate: we first create comparable counts of non-Hispanics enrolled in private school;[4] we next aggregate these counts to consistent secondary district boundaries; and finally, we divide by the aggregated district's 5 to 19 year old population.

From the 1970 data, we also collected information on the distribution of foreign-born Mexicans across school districts, used in construction of the instrument, as well as non-Hispanic median family income and the share of 16 to 17 year olds not enrolled in school.

---

[1] For California residents, the Spanish Heritage population includes "persons of Spanish language or persons not of Spanish language but of Spanish surname identified by matching with a list of about 8,000 such names."

[2] To avoid disclosure, cell values are rounded; generally, this rounding is to the nearest 5, or to 4, when the population count is under 5. On a few occasions, rounding leads to (small) negative values.

[3] In 1970, counts of residents by gender were originally reported for the total population and for the "Spanish Heritage" (hereafter referred to as Hispanic) population in detailed age bins (Table 17). In 2000, counts of residents by age and gender were reported for the total population (Table P8 for Total – Population and Households (TT)) and for the Hispanic/Latino population (Table 145H for TT).

[4] The 1970 data reports counts of residents aged 3 to 34 enrolled in private school, by level (kindergarten, elementary, and secondary), for the total population and for the Hispanic population (Table 38). The 2000 data report counts of residents in private school, by gender, separately for all children and for Hispanic/Latino children either enrolled in or of age to be enrolled in the grades served by the district (Tables P8 and 145H for Children (CO): Relevant Children – Enrolled Private). For consistency, we limit counts of non-Hispanic private school enrollees in 1970 to the grade levels served by the district, and aggregate the 2000 figures across age and gender.

2. 1976 and 2000 Office for Civil Rights Data

For 1976, school district level data on the number of low-English students, by race/ethnicity, were drawn from the Fall 1976 Elementary and Secondary School Civil Rights Survey (Office for Civil Rights (OCR) 1978). The 1976 OCR survey covered all school districts in the United States. For 2000, school district level data on the number of low-English students by ethnicity were drawn from the 2000 Elementary and Secondary School Civil Rights Compliance Report District Survey (OCR 2000). The 2000 OCR survey covered all school districts in the United States, with tabulations rounded to the nearest 5, to avoid disclosure.

The original data give counts of "pupils whose primary language is other than English" in total and by race/ethnicity. Our treatment variable is constructed using the number of Hispanics (of all races) with this designation and total enrollment. We drop districts for which either of the following held in either 1976 or 2000: (1) the sum of non-low-English enrollment by race was more than 10 percent above or below reported non-low-English enrollment; or (2) the sum of enrollment by race was more than 10 percent above or below reported enrollment.

3. Other data sources

We use a number of data sources to construct additional district-level covariates. Data on per-pupil property tax revenues, per-pupil total expenditures, and the pupil-teacher ratio for 1971-72 are from *Census of Governments, 1972: Government Employment and Finance Files* (Bureau of the Census 1972). The indicator for having an above-average number of public schools as of 1972 given land area and 1976 enrollment (from OCR) was calculated using counts of public schools in 1972 reported in *Elementary and Secondary General Information System (ELSEGIS): Public School District Universe Data, 1972-1973* (NCES 1973) and land area data information in GeoLytics Neighborhood Change Data Base. Information on center city status was also drawn from NCES (1973). All raw data were aggregated to consistent secondary district boundaries for 1970 to 2000 prior to construction of the variables used in the analysis.

Tract-level data on non-Hispanic population for 1960 are from NHGIS (Minnesota Population Center 2004).

B. School District Geography

We identify school district reorganizations using data from the Elementary and Secondary Education General Information System and the Common Core of Data Public Agency Universe and internet searches. We drop "aggregated" districts involved in reorganizations over the period if any of the component districts are not present in years they should be or, in a few cases, if we were not able to ascertain the nature of the reorganization that occurred.

C. Matching of 1960 Tracts to School Districts

In many cases 1960 and 1970 tract boundaries were identical. In the cases where they were not, we used published Census tabulations of the correspondence between 1960 and 1970 tracts

(Table B in Bureau of the Census 1972) to construct collections of tracts that could be used to identify the smallest possible identical geographic regions in each census. For example, if a 1960 tract was split into two pieces, we would use that tract in the 1960 data and the aggregate of the two corresponding tracts in the 1970 data. In some cases the overlap between tracts was more complex than this example, but it was almost always possible to construct an exact match by aggregating enough tracts in both years.[5] We then used the *School District Geographic Reference File, 1969-1970* (Bureau of the Census 1970) to determine the fraction of each tract aggregate's total population inside the borders of each school district in 1970. We apportioned non-Hispanics in each "tract aggregate" to school districts with these weights – which were mostly 0 or 1 – in 1960.

D. References

Bureau of the Census. 1970. *School District Geographic Reference File, 1969-1970* [Computer file]. ICPSR version. Washington, DC: U.S. Dept of Commerce, Bureau of the Census [producer]. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 2002.

Bureau of the Census. 1972. *Census of Governments, 1972: Government Employment and Finance Files* [Computer file]. ICPSR ed. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [producer and distributor], 1993.

Bureau of the Census. 1972. "Census of Population and Housing: 1970 Census Tracts Final Report PHC(1)" Washington, DC: U.S. Government Printing Office.

Bureau of the Census. 1981. *Survey of Income and Education, 1976: Rectangular File* [Computer file]. Conducted by U.S. Dept. of Commerce, Bureau of the Census, and University of Michigan, National Chicano Research Network. ICPSR07919-v1. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [producer and distributor].

Minnesota Population Center. 2004. National Historical Geographic Information System: Pre-release Version 0.1. Minneapolis, MN: University of Minnesota. http://www.nhgis.org/ (accessed March 28, 2010).

National Center for Education Statistics. 1970. *User's Manual for 1970 Census Fourth Count (Population): School District Data Tape* [Computer file]. ICPSR version. Washington, DC: United States Department of Education, National Center for Education Statistics [producer]. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 2004. doi:10.3886/ICPSR03525

National Center for Education Statistics. 1973. *Elementary and Secondary General Information System (ELSEGIS): Public School District Universe Data, 1972-1973* [Computer file]. ICPSR version. Washington, DC: United States Department of

[5] The only exception to this was there were a handful of 1970 tracts or parts of tracts on the edge of metro areas that were untracted in 1960.

Education, National Center for Education Statistics [producer]. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 2000.

National Center for Education Statistics. 2000. *Census 2000 School District Tabulation (STP2) Data Download*. Washington, DC: United States Department of Education, Institute of Education Sciences. http://nces.ed.gov/surveys/sdds/downloadmain.asp (accessed June 18, 2010).

Office for Civil Rights. 1978. *Fall 1976 Elementary and Secondary School Civil Rights Survey*. Washington, DC: United States Department of Health, Education and Welfare [producer]. Los Angeles, CA: Sarah Reber [distributor, computer file], 2006. http://l1.ccpr.ucla.edu/OCR/ocr.htm (accessed October 30, 2009).

Office for Civil Rights. 2000. *Elementary and Secondary School Survey 2000*. Washington, DC: United States Department of Education. http://www.ed.gov/about/offices/list/ocr/data.html (accessed October 31, 2009),

Online Appendix B: Derivation of Estimation Equation

A reduced-form, linear expression of the model in Section I is given by:

$$Y^{j}_{n,d,m,t} = \gamma_{d,t} + \beta_j k_{d,t} + \eta^{j}_{n,d,m,t},$$

where $Y^{j}_{n,d,m,t}$ = 1 if non-Hispanic household *n* of type *j* resides in school district *d*, which is located in metropolitan area *m*, at time *t*.[6] This decision is a function of the share of public school enrollees who are low-English Hispanics in *d* at time *t*, $k_{d,t}$, the response to which varies by household type, with *j*=1 for non-Hispanic households with children (the treatment group), and *j*=0 for the comparison group. This decision is also a function of other district characteristics at time *t* not differentially valued across types, captured by $\gamma_{d,t}$. Conceptually, $\gamma_{d,t}$ captures reactions to all district amenities that the two household types hold in common (including the other effects of *i* in the model in Section I). The parameter of interest is $\theta \equiv \beta_1 - \beta_0$, the difference across household types in the sensitivity of location decisions to *k*.

We are not able to estimate this equation given a lack of household-level data with sufficient geographic detail. To get equation (2), we begin by summing this across all non-Hispanic households *n* of type *j* in metro area *m* at time *t*:

(B1) $$N^{j}_{d,m,t} = N^{j}_{m,t}\left(\gamma_{d,t} + \beta_j k_{d,t}\right) + \sum \eta^{j}_{n,d,m,t}$$

The ideal household-level model at the beginning of Section II, (2), divides through (B1) by $N^{j}_{m,t}$ and then differences over time and across types to eliminate the effects of all common factors affecting location at a point in time (the $\gamma_{d,t}$):

$$\Delta\left(N^{1}_{d,m} / N^{1}_{m}\right) - \Delta\left(N^{0}_{d,m} / N^{0}_{m}\right) = \theta \Delta k_d + \varepsilon_{d,m}$$

As noted, public use data with counts of households are insufficiently detailed to estimate this equation directly with our comparison group, childless households with a householder of parenting age (under age 50). Due to this data constraint, we instead use population data to estimate (3), in which the dependent variable is the difference between the district's 0 to 19 year old and 20 to 49 year old metropolitan shares. As a result of including parents in the comparison group, we noted, our estimates understate the parameter of interest.

To see this, let $\tau^{j,a}_{m,t}$ represent the average number of individuals in age group *a* (=0-19 or 20-49) per household of type *j* in *m* at *t*. Multiplying (B1) by $\tau^{j,a}_{m,t}$ and summing across household types

[6] *Y* in equation (2) might also be interpreted as the latent propensity to live in district *d*, with the probability of living in the district a non-linear transformation of it. If this transformation is a logistic CDF, then equation (2) would be the same except that the share of type *j* households in *d* will be replaced with the log odds that a type *j* household is in *d* (equal to ln(share/[1-share])), i.e., the logit model.

within age group generates a model for the (approximate) number of individuals of age *a* in district *d*:

$$N_{d,m,t}^{0,a} + N_{d,m,t}^{1,a} = N_{m,t}^{0,a}\left(\gamma_{d,t} + \beta_0 k_{d,m,t}\right) + N_{m,t}^{1,a}\left(\gamma_{d,t} + \beta_1 k_{d,m,t}\right) + \sum\left(\eta_{n,d,m,t}^{0,a} + \eta_{n,d,m,t}^{1,a}\right),$$

where $N_{d,m,t}^{j,a} \equiv \tau_{m,t}^{j,a} N_{d,m,t}^{j}$ and $N_{m,t}^{j,a} \equiv \tau_{m,t}^{j,a} N_{m,t}^{j}$. Letting $N_{d,m,t}^{a} \equiv N_{d,m,t}^{0,a} + N_{d,m,t}^{1,a}$ and $N_{m,t}^{a} \equiv N_{m,t}^{0,a} + N_{m,t}^{1,a}$, and noting that population aged *a* in *m* at *t* is $N_{m,t}^{a} = \sum_{d \in m} N_{d,m,t}^{a}$, this model can be rewritten as

(B2) $$N_{d,m,t}^{a} / N_{m,t}^{a} = \gamma_{d,t} + \left(\beta_0 + \varphi_{m,t}^{a}\left(\beta_1 - \beta_0\right)\right)k_{d,m,t} + u_{d,m,t}^{a},$$

where $\varphi_{m,t}^{a} \equiv N_{m,t}^{1,a} / N_{m,t}^{a}$ is the fraction of individuals aged *a* in *m* at *t* who are in households with any 0-19 year olds. Thus, by definition, $\varphi_{m,t}^{0-19} = 1$ for all *m* and *t*. To get to our estimation equation, we simplify further by assuming that there is no variation over time and across metropolitan areas in this parameter for 20-49 year olds, or that $\varphi_{m,t}^{20-49} = \varphi^{20-49}$. Differencing over time within age groups, then across age groups then yields:

(B3) $$\Delta\left(N_{d,m}^{0-19} / N_{m}^{0-19}\right) - \Delta\left(N_{d,m}^{20-49} / N_{m}^{20-49}\right) = \left(1 - \varphi^{20-49}\right)\theta\,\Delta k_d + \left(\Delta u_d^{0-19} - \Delta u_d^{20-49}\right).$$

The parameter of interest remains θ, but as (B3) shows, the slope parameter we actually estimate is $\widetilde{\theta} \approx \left(1 - \varphi^{20-49}\right)\theta$. So our estimate is attenuated by a factor roughly proportional to the share of 20 to 49 year olds in households with children.[7]

An alternative approach would have been to estimate (B2) more directly by interacting the treatment with time x metro varying estimates of $\varphi_{m,t}^{20-49}$, rather than assuming away the variation in $\varphi_{m,t}^{20-49}$. Our view is that our simpler approach is more transparent, and therefore more convincing. Nevertheless, we have estimated (a transformed version of) (B2), described here.

A more practical challenge in directly estimating (B2) is that metropolitan-specific estimates of the fraction of families with kids are not available in the initial period (because the 1970 Census has less geographic detail than the 2000 Census). However, differencing (B2) between age groups and over time reveals that a more general estimation strategy can be employed without this information:

[7] The logit and growth specifications in Table 5 are also biased by approximately the same factor, and so their coefficients are also divided by this in the calculation of household-level displacement rates. Converting the estimate in row (d) of Table 5 to a household-level displacement rate is not as simple, but also depends on this factor. (This and other details of our displacement calculations are available on request.)

$$\Delta\left(N_{dm}^{0-19}/N_{m}^{0-19}\right)-\Delta\left(N_{dm}^{20-49}/N_{m}^{20-49}\right)=\left(1-\varphi_{m,2000}^{20-49}\right)\theta\,\Delta k_{d}-\Delta\varphi_{m}^{20-49}\theta k_{d,0}+\left(\Delta u_{d}^{0-19}-\Delta u_{d}^{20-49}\right)$$

This expression differs from (B3) in two ways. First, the $\left(1-\varphi^{20-49}\right)$ interacted with $\theta\,\Delta k_d$ varies across metropolitan areas according to its end year (2000) values. Second, a term capturing the impact of changes in φ_m^{20-49} is in the equation, i.e. $\Delta\varphi_m^{20-49}k_{d,0}$. Note that this term is an omitted variable in the simplified approach we take, but we cannot measure or control for it directly. However, it can be absorbed by allowing initial low-English Hispanic share to have metro-specific effects, as in:

(B4) $$\Delta\left(N_{dm}^{0-19}/N_{m}^{0-19}\right)-\Delta\left(N_{dm}^{20-49}/N_{m}^{20-49}\right)=\left(1-\varphi_{m,2000}^{20-49}\right)\theta\,\Delta k_{d}+\gamma_{m}k_{d,0}+\xi_{dm}.$$

$\gamma_m k_{d,0}$ captures the effects of $\Delta\varphi_m^{20-49}k_{d,0}$.

We have estimated (B4). The $\gamma_m k_{d,0}$ control makes our estimate of θ larger in magnitude. Allowing for metro-specific variation in φ_m^{20-49} makes the estimate of θ a little larger still. The basic reason is that areas which received more immigrants relative to their population, such as in the Central Valley, tended to have a higher share of families with kids. Thus, the main approach we take in this paper is conservative: by assigning the average fertility rate in the state to these high treatment areas, we underadjust our coefficient estimates. The bottom line is that if we could actually obtain appropriate household count data to employ our estimation strategy, the estimated household displacement rates would likely be larger in magnitude than those we report in Tables 4 and 5.