Online Data Appendix

In this data appendix, I include several sections to describe in detail the construction of the data used in this analysis as well as any peculiarities of these data.

Data from the Anuarios

Most of the data used in this analysis come from a series of statistical yearbooks from Mexico entitled, *Anuarios estadisticos de los Estados Unidos Mexicanos*. The data that I transcribe from these sources are at the state by year level and include the number of braceros leaving a state in a given year, primary school enrollments (urban, rural, and total), post-primary school enrollments (male, female, and total), primary schools, and state education spending. I transcribed these data for twenty four years (1942 to 1965) for the thirty two states and territories in Mexico. This should yield a sample of 768 observations. However, given that some of the data are missing for some of the states and years throughout the period of the study, the sample sizes in the reported regressions vary. I created Data Appendix Table 1 to explicitly describe which observations are missing for all of the variables of interest in the main analysis.

As a result of these missing values, the sample sizes of different regressions will vary throughout the paper. For example, the first stage regression in Table 4 has a sample size of 620, all of the non-missing entries for the natural log of braceros (i.e., 736 potential observations minus 116 missing values). One can similarly use the missing observations for both treatment and outcomes that I provide in Data Appendix Table 1 to trace the differences in sample sizes across the main regressions in the paper. I explicitly describe these in Data Appendix Table 2.

Data from IPUMS

In order to explore heterogeneity in my estimates by age and by sex, I use the 1970 sample of the Mexican census from IPUMS. I use data about birthplace and age in order to create an

estimate of age- and sex-specific populations for each state and each year of the study period. I further use information about completed years of schooling to generate estimates of the proportion of each age- and sex-specific population is in school for each state and year of the study period. Firstly, I must drop some cases in order to produce these estimates. I drop the 347 cases where age is missing, the 1 case where the schooling variable is missing, 26 cases where schooling is given as "some primary," and the 1,675 cases where the individual was not Mexican-born. Secondly, I have to make some strict assumptions in order to estimate state and year specific populations and enrollment rates by age and sex. I assume that state in which one is born is the state in which they stay until they are greater than 18 years of age (i.e., no permanent internal or international migration for children). I also assume that all schooling begins at age six and continues consecutively without interruption until the child leaves school. The final assumption that I make is that one's birth year is the year of the census (1970) minus their age plus one year. For example, if a person is 19 years old in the 1970 census, I code their birth year as 1970 – (19 +1) = 1950. Given that the census date was January 28, 1970, this will pin down the correct birth year for all people born after January 28th (the vast majority of the year), but will underestimate it by one year for those born between January 1st and January 28th. Using these assumptions, I can construct my estimates. I understand there are many problems associated with these assumptions (I discuss them explicitly in the text) and that is why I do not use these measures for the main analysis in the paper.

I will use an example here to illustrate how I construct the state-by-year-by-age-by-sex population and enrollment estimates from the IPUMS cross-section. Suppose there is a female who has an birth year of 1950 in the 1970 census, who was born in Chihuahua, and who has reported two years of schooling. I count that individual as a six year old female in Chihuahua in 1956, a seven year old female in Chihuahua in 1957, an eight year old female in Chihuahua in 1958, and

so on and so on. Then, I count that individual as a six year old female in school in Chihuahua in 1956, a seven year old female in school in Chihuahua in 1957, an eight year old female not in school in Chihuahua in 1958, and so on and so on. After I do this for ever individual in the IPUMS sample, I can divide the number in school by the total number for each state-by-year-by-age-by-sex cell to generate the enrollment rates that I use in the paper.

Calculating Distance to Recruitment Centers

The instrumental variable in this paper relies on measuring the distance to the nearest recruitment center from a given state in a given year. There are two issues that arise in constructing such a variable. First of all, a state is a polygon, not one particular point or set or coordinates, and so a point needs to be chosen within a state to use to measure the distance to the coordinates of the recruitment centers. I use two different ways of doing this. One way is to measure the distance between recruitment centers and the centroid of the state; a method that I refer to as the centroid method in the paper. Another way is to randomly select 200 different points in each state, measure the distance between those points and the recruitment centers, and then take the average over the 200 points; a method that I refer to as the point method in the paper. The second problem that I must overcome involves determining how long a state needs to be exposed to the recruitment center in a given year for it to be included in the determination of the center at the closest difference. I also use two different ways of doing this. One way is to count a recruitment center as the closest center to the state if it was the shortest distance to the state (as determined by one of the methods above) for a majority of the year (i.e., six months or more); a method that I refer to as the majority method in the paper. Another way is to count a recruitment center as the closest center to the state if it was the shortest distance to the state (as determined by one of the methods above) for any part of the year; a method that I refer to as the partial method in the paper. These two

methods of calculating distance and two methods of determining length of time for treatment yield four different possibilities that I refer to in the paper as centroid-majority, centroid-partial, point-majority, and point-partial. I use centroid-majority for the main part of the paper as it is the most conservative approach, although I test the robustness of the results to the other measures in the Online Appendix.

Data Appendix Table 1 – Explanation of Missing Observations

Variable	Missing Observations
Braceros	All states (1955-1957) and Baja California Sur (1958-1964) for a total of 103 missing observations. There are also thirteen zero values that become missing/undefined when the natural log is used. These include Baja California Sur (1942, 1944, 1945, 1946, and 1948), Quintana Roo (1942, 1944, 1946-1948, 1953, and 1963), and Sinaloa (1946).
Primary School Enrollment	All states (1961) for a total of 32 missing observations. There are also ten zero values for rural primary school enrollments that become missing/undefined when the natural log is used. These include DF (1942-1949 and 1964-1965).
Post-Primary School Enrollment	All states (1942-1949) for a total of 256 missing observations. There is one zero value for post-primary enrollments (men, women, and total) that become missing/undefined when the natural log is used. These include Quintana Roo (1959).
State Education Spending	All states (1963), Baja California (1946-1950), Baja California Sur (1946-1959 and 1965), DF (1946-1948), Guanajuato (1965), Guerrero (1961), Puebla (1965), Quintana Roo (1946-1959, 1962, and 1964), and Sinaloa (1954) for a total of 75 missing observations. There are also eight zero values that become missing/undefined when the natural log is used. These include Baja California (1944-1945), DF (1942 and 1944), and Quintana Roo (1942-1945).

Source: Author's calculations of sample sizes.

Data Appendix Table 2 – Sample Sizes by Regression

Regression	Sample Size	Explanation
Table 4 Column 1; Table 7 Columns 1-13	620	736 total potential observations minus 116 missing bracero observations.
Table 5 Columns 1 and 3; Table 6 Columns 1 and 3	589	736 total potential observations minus 116 missing bracero observations minus 31 additional missing primary school enrollment observations.
Table 5 Column 2; Table 6 Column 2	580	736 total potential observations minus 116 missing bracero observations minus 31 additional missing primary school enrollment observations minus 9 additional missing rural primary school enrollment observations.
Table 5 Column 4; Table 6 Column 4	589	736 total potential observations minus 116 missing bracero observations minus 31 additional observations because treatment is not defined in 1942 for period t-1.
Table 5 Column 5; Table 6 Column 5	529	736 total potential observations minus 116 missing bracero observations minus 31 additional observations because treatment is not defined in 1942 for period t-1 minus 60 additional missing state education spending observations.
Table 5 Columns 6-8; Table 6 Columns 6-8; Table 8 Columns 6-8	374	736 total potential observations minus 116 missing bracero observations minus 246 additional missing post-primary enrollment observations.
Table 8 Columns 1-5	Various	Sample sizes correspond to the same columns in Tables 5 and 6 minus one more observation for missing information on the percentage of PRI votes in the last election for Baja California Sur in 1943.

Source: Author's calculations of sample sizes.