J Econ Growth (2006) 12:101–158
DOI 10.1007/s10887-007-9016-0

Education and income of the states of the United States: 1840–2000

Chad Turner · Robert Tamura · Sean E. Mulholland · Scott Baier

Published online: 13 June 2007

Abstract This article introduces original annual average years of schooling measures for each state from 1840 to 2000. Our methodology results in state estimates similar to those reported in the United States Census from 2000 back to 1940 and national, turn of the century estimates strikingly close to those presented by Schultz (Schultz, T. (1961). In N. B. Henry (Ed.), *Social forces influencing American education*. Chicago: University of Chicago Press.) and Fishlow (Fishlow, A. (1966). In H. Rosovsky (Ed.), *Industrialization in two systems*. John Wiley & Sons). To further determine the validity of our state schooling estimates, we first combine original data on real state per worker output with existing data to provide a more comprehensive series of real state output per worker from 1840 to 2000. We then estimate aggregate Mincerian earnings regressions and discover that the return to a year of schooling for the average individual in a state ranges from 11% to 15%. This range is robust to various time periods, various estimation methods, various assumptions about the endogeneity of schooling and is in line with the body of evidence from the labor literature.

Keywords State years of schooling · State real output per worker

All views expressed here are the authors' and not necessarily those of the Federal Reserve Bank of Atlanta or the Federal Reserve System.

C. Turner
Nicholls State University, Thibodaux, LA, USA

R. Tamura (✉) · S. Baier
Clemson University, Clemson, SC, USA

R. Tamura · S. Baier
Atlanta Federal Reserve Bank, Atlanta, GA, USA

S. E. Mulholland
Mercer University, Macon, GA, USA

1 Introduction

This paper makes two fundamental contributions: (1) it introduces original annual years of schooling and average years of experience measures in the labor force for each of the states of the United States, generally from 1840 to 2000, and (2) it constructs original real state per worker output estimates for 1850, 1860, 1870, 1890 and 1910, and combines them with existing data for 1840, 1880, 1900 and 1920 and 1929 through 2000. Furthermore, it captures the educational choices made by individuals (aggregated to the state level) over much of the history of the United States. To construct these measures we make use of data from the decennial censuses of the United States, Richard Easterlin's work on state income, Thomas Weiss's state estimates of the labor force in the nineteenth century, *Historical Statistics of the United States: Colonial Times to 1970* (US Department of Commerce 1975) as well as information contained in annual *Statistical Abstracts of the United States* to produce these estimates.[1] Even with these numerous data sources we are required to make various assumptions that, although not always ideal, are a result of the dearth of information available at the state level over much of nineteenth and early 20th century.[2]

To check the validity of our state-level estimates we estimate the relationship between the level of state education and income. We estimate the return to a year of schooling for the average individual in a state ranges from 11% to 15%. This range is robust to various time periods and various estimation methods. We view this work as complementary to the work of Mulligan and Sala-i-Martin (1997, 2000).[3] We also document the long-term enrollment trends in primary, secondary, and tertiary schooling as well as the patterns of income growth across census regions. We show both within region and across region convergence.

The remainder of the paper is organized as follows: the next section provides the accounting framework for calculating average years of schooling by state. We present in graphical and tabular form the results of these calculations by census region. Section 3 presents our measures of state output per worker. Section 4 contains our estimates of returns to schooling and returns to potential experience. We find that OLS estimates are quite robust to alternative specifications, and that a year of schooling returns about 14% to an individual in additional productivity. Section 5 concludes and describes broader implications and future work.

2 Education in the states

We use a perpetual inventory method, employed by Barro and Lee (1993) and Baier, Dwyer, and Tamura (2006), to construct average years of schooling in the labor force for each state. Because we are interested in the relationship between human capital and output per worker, it is more appropriate to calculate the average years of schooling in the labor force instead

[1] Data covering a large number of states (28) is first available in 1840. Before 1840, we are aware of enrollment data for nine states: Maine, New Hampshire, Connecticut, Rhode Island, Massachusetts, New York, South Carolina, Virginia, and Kentucky. For a greater discussion of schooling in the first half of the nineteenth century see Fishlow (1966).

[2] We also admit that the accuracy of these enrollment data have been questioned by previous analyses. The American Statistical Association offered an official critique of the 1840 Census and found errors in the collection of common school data (Senate Document No. 5, 28th Congress, 2nd Session). We are comforted however, by Fishlow's 1966 conclusion that "for most purposes [the Census statistics] seem to suffice in their present form."

[3] Mulligan and Sala-i-Martin (1997, 2000) construct two different state level human capital measures for the census years 1940–1990, inclusive. Our years of schooling measure is highly correlated with theirs, averaging approximately 0.8. See Appendix D for more detail.

of the average years of schooling of all state residents.[4,5,6] Enrollment data from United States Censuses, Digests of Education Statistics and Statistical Abstracts of the United States present the number of students enrolled in one of three educational categories: primary, secondary, and college.[7] In order to calculate the average years of schooling in the work force, our methodology must account for:

1. the number of school age children;
2. the number of new labor force participants, I_t^i, and their education level;
3. the departure rate of workers from the workforce, δ_t^i;
4. the interstate migration of students post education;
5. and the impact of foreign-educated immigrants.

We assume that there are four categories of workers: those with no schooling (none); those exposed to primary schooling and no more (primary); those exposed to secondary schooling and no more (secondary); and those with exposure to higher education (college). Suppressing the state subscript, H_t^i is the number of workers in the labor force in year t in education category i. The perpetual inventory method produces the following law of motion:

$$H_{t+1}^i = H_t^i \left(1 - \delta_t^i\right) + I_t^i, \ i = \text{none, primary, secondary, college} \tag{1}$$

where δ_t^i is the departure rate from the labor force between year t and $t+1$ and I_t^i is the gross flow of new workers into the labor force from education category i.

In order to get estimates of the flows into each education category, we use the following information:

$$I_t^{\text{college}} = \frac{r_t^{\text{college}} \Theta_t lfpr^{\text{college}} \ell[18-24]_t}{7} \tag{2}$$

$$I_t^{\text{secondary}} = \frac{\left(r_t^{\text{secondary}} - r_t^{\text{college}} \Theta_t\right) lfpr^{\text{secondary}} \ell[14-17]_t}{4} \tag{3}$$

[4] Additional details on the derivation and the data sources are furnished in Appendix B.

[5] Ideally we would use information to produce average years of schooling for men and women separately in the labor force, however, enrollment information by sex is not consistently available. However Series H 433–441, page 370 of *Historical Statistics of the United States: Colonial Times to 1970*, indicates that there was little difference in enrollment rates of men and women:

Sex	1850	1860	1870	1880
Male	49.6	52.6	49.8	59.2
Female	44.8	48.5	46.9	56.5

From 1890 onward, differences in enrollment rates were less than one percentage point. We acknowledge that our calculations implicitly assume the labor force participation rate is common across men and women.

[6] We are unable to account for changes in the labor force participation rates by educational category because we do not have data on labor force participation by education category prior to 1960.

[7] See Appendix A for additional information on enrollment rates by educational category. The information from the various issues of the Statistical Abstracts of the United States are the identical data contained in the Annual Reports of the Commissioner of Education of the US Interior Department. A summary volume of the latter is available at http://nces.ed.gov/pubs93/93442.pdf, entitled *120 Years of American Education: A Statistical Portrait*.

$$I_t^{\text{primary}} = \frac{\left(r_t^{\text{primary}} - r_t^{\text{secondary}}\right) lfpr^{\text{primary}} \ell[5-13]_t}{9} \tag{4}$$

$$I_t^{\text{none}} = \frac{\left(1 - r_t^{\text{primary}}\right) lfpr^{\text{primary}} \ell[5-13]_t}{9} \tag{5}$$

where in year t, r_t^i is the enrollment rate in education category i, $lfpr_t^i$ is the labor force participation rate for each educational category, and $\ell[i-j]_t$ is the population in age category $[i-j]$, inclusive.[8] We assume that population within each age category is uniformly distributed and that education enrollment rates are constant across ages within the primary and secondary education categories. The constant Θ_t is an adjustment for the fact that, because there is high rate of attrition in the early part of higher education, assuming a uniform enrollment rate across ages will understate the true inflow into the higher educational category.[9]

Although values of are δ_i^t are not directly available, we are able to calculate three different departure rates: one for college workers, $\delta_t^{\text{college}}$, one for secondary workers, $\delta_t^{\text{secondary}}$, and one for all other workers, $\delta_t^{\text{primary}}$ using the following solution:[10,11]

First, we assume that workers with some college exposure do not disappear at a calculated rate, but only after 45 years of employment. Thus for college exposed workers, the law of motion becomes:

$$H_{t+1}^{\text{college}} = H_t^{\text{college}} - I_{t-45}^{\text{college}} + I_t^{\text{college}} \tag{6}$$

We let h_{t+1}^i represent the share of the labor force exposed to educational category i. Dividing the law of motion equation by the labor force in period $t+1$ for the higher educational category provides:

$$h_{t+1}^{\text{college}} = h_t^{\text{college}} \frac{L_t}{L_{t+1}} - \frac{I_{t-45}^{\text{college}}}{L_{t+1}} + \frac{I_t^{\text{college}}}{L_{t+1}} \tag{7}$$

[8] For labor force participation rates by educational attainment we used data from the 1940–2000 censuses. We use .91, .82 and .60 for $lfpr^{\text{college}}$, $lfpr^{\text{secondary}}$, and $lfpr^i$, $i=$primary, none, respectively. We used these labor force participation rates for the entire 1840–2000 period. While it may seem strange to use a constant labor force participation rate, in 1840 the labor force participation rate for 14–65 year old individuals was 51% and in 1900 the labor force participation rate for this same category was 57%. Since little information is available by educational category and the majority of our labor force is either without education or with only primary education in this early period, holding labor force participation rates constant over time across education categories is reasonable.

[9] Since our calculations of the inflow to all categories are equal to the total enrollment across all ages in the category divided by the total population across all ages in the category, they implicitly assume the enrollment rate is constant across ages within each education category. To the extent that this assumption is erroneous, the true inflow in to the category will be understated. While this assumption is implicit in our calculations for inflows into all educational categories, it is most problematic where there is a high attrition rate between ages. Because attrition rates are highest between the first and second years of higher education, we multiply the measured inflow into the higher education category by a factor denoted Θ_t. We allow Θ_t to take different values in eight subperiods: 1840–1940 and decade specific values from 1940 to 2000. Within the 1840–1940 subperiod, we assume Θ_t is time invariant and does not vary across states. Within the 1940–2000 period, we assume theta is decade specific for each state. For additional details, see Appendix B.

[10] We use a common departure rate for the primary and none educational categories, which we denote $\delta_t^{\text{primary}}$.

[11] The creation of a separate departure rate for college workers is a motivated by the fact that a common departure rate for all education categories produces a share of workers exposed to higher education significantly below the value reported in the census in 2000. After making this adjustment, we further find that a departure rate common to the remaining classes (secondary, elementary and none) produces some states where the share of workers exposed to elementary schooling is less than zero. As a result, we also allow for a separate departure rate for those workers exposed to secondary schooling.

For the very early years, $I_{t-45}^{\text{college}}$ is approximated using the first observed measure of higher education enrollment rates in t.[12] Once enough years have past, we use our own calculations for $I_{t-45}^{\text{college}}$.

Second, for workers exposed to secondary schooling, we choose $\delta_t^{\text{secondary}}$ by utilizing decennial census data on the share of workers exposed to secondary education from 1940 to 2000. Given the structure of our laws of motion and inflow calculations, we choose the value of $\delta_t^{\text{secondary}}$ that results in the closest match of the evolution of $h_t^{\text{secondary}}$ to that of the corresponding census data from 1940 to 2000.[13] For values, see Appendix B. The result is:

$$h_{t+1}^{\text{secondary}} = h_t^{\text{secondary}} \frac{L_t}{L_{t+1}} \left(1 - \delta_t^{\text{secondary}}\right) + \frac{I_t^{\text{secondary}}}{L_{t+1}} \tag{8}$$

Third, though we are unable to calculate the departure rate for the remaining educational classes directly, we can isolate $\delta_t^{\text{primary}}$ using the following identity:

$$L_{t+1} = H_{t+1}^{\text{college}} + H_{t+1}^{\text{secondary}} + H_{t+1}^{\text{primary}} + H_{t+1}^{\text{none}} \tag{9}$$

Dividing through by L_{t+1} and then substituting using (1) for the primary and none categories yields:

$$\begin{aligned} 1 = {} & \frac{H_{t+1}^{\text{college}}}{L_{t+1}} + \frac{H_{t+1}^{\text{secondary}}}{L_{t+1}} + \frac{H_t^{\text{primary}}\left(1 - \delta_t^{\text{primary}}\right) + I_t^{\text{primary}}}{L_{t+1}} \\ & + \frac{H_t^{\text{none}}\left(1 - \delta_t^{\text{primary}}\right) + I_t^{\text{none}}}{L_{t+1}} \end{aligned} \tag{10}$$

$$\begin{aligned} 1 - h_{t+1}^{\text{college}} - h_{t+1}^{\text{secondary}} = {} & \left(h_t^{\text{primary}} + h_t^{\text{none}}\right) \frac{L_t}{L_{t+1}} \left(1 - \delta_t^{\text{primary}}\right) \\ & + \frac{I_t^{\text{primary}} + I_t^{\text{none}}}{L_{t+1}} \end{aligned} \tag{11}$$

We then isolate our estimate of the $\frac{L_t}{L_{t+1}}\left(1 - \delta_t^{\text{primary}}\right)$ term:

$$\frac{L_t}{L_{t+1}}\left(1 - \delta_t^{\text{primary}}\right) = \frac{1 - h_{t+1}^{\text{college}} - h_{t+1}^{\text{secondary}} - \left(\frac{I_t^{\text{primary}} + I_t^{\text{none}}}{L_{t+1}}\right)}{\left(h_t^{\text{primary}} + h_t^{\text{none}}\right)} \tag{12}$$

Thus for the share of labor force with primary schooling exposure we return to (1), divide by L_{t+1} and produce:

$$h_{t+1}^{\text{primary}} = h_t^{\text{primary}} \frac{L_t}{L_{t+1}} \left(1 - \delta_t^{\text{primary}}\right) + \frac{I_t^{\text{primary}}}{L_{t+1}} \tag{13}$$

[12] This is not much of an issue in the early years because higher education enrollments are near zero. Further details are discussed in Appendix B.

[13] We simply utilize our methodology for each value of $\delta_t^{\text{secondary}}$ across a grid in increments of 0.0001. We select the value of $\delta_t^{\text{secondary}}$ for each state and for each decade that most closely matches our calculated data to the census data.

and then use the following adding up restriction for the share of the labor force with no educational exposure:

$$h_{t+1}^{\text{none}} = 1 - h_{t+1}^{\text{college}} - h_{t+1}^{\text{secondary}} - h_{t+1}^{\text{primary}}.^{14} \tag{14}$$

We use information from the 1940–2000 Censuses to get estimates for the expected number of years of schooling completed, conditional on being in each education category for each state. These expected years of schooling by category are represented by $yrs_{it}^{\text{college}}$, $yrs_{it}^{\text{secondary}}$ and $yrs_{it}^{\text{primary}}$. For the intervening years we log linearly interpolate. Initial values for $yrs_{it}^{\text{college}}$, $yrs_{it}^{\text{secondary}}$ and $yrs_{it}^{\text{primary}}$ are set at 14, 10 and 4, respectively, in the year that data becomes available for each state.[15] We then log linearly interpolate from these initial values to the 1940 value. Thus for state i we calculate average years of schooling in the labor force as:

$$\widehat{E}_{it} = h_{it}^{\text{college}} yrs_{it}^{\text{college}} + h_{it}^{\text{secondary}} yrs_{it}^{\text{secondary}} + h_{it}^{\text{primary}} yrs_{it}^{\text{primary}} \tag{15}$$

To account for interstate migration, we adjust our years of schooling measure by residents' state of birth reported in the 1850 through 2000 Censuses. We assume that all education is undertaken in an individual's state of birth and that all current migrants are educationally representative of their birth state. Due to data limitations, we can not allow for differential rates of migration by educational attainment.[16] Let $\widehat{E}_{jt}$ be the years of schooling at time t for those born in state j. Our estimate of years of schooling in state i therefore is:

$$E_{it} = \sum_{j=1}^{52} S_{ijt} \widehat{E}_{jt} \tag{16}$$

where S_{ijt} is the share of state i residents in year t that were born and educated in state j.[17] There are 52 categories where workers could have received their education: 50 states, the District of Columbia, and the foreign born. For the foreign born we assume that the individuals come from the kth percentile of the primary, secondary and higher education distributions. We use the information from each of the 1940–2000 Censuses to determine the best fitting kth percentile for each state and census year in order to match the state's average years of

[14] There are occasions when $h_t^{\text{none}} < 0$. In these instances, we set $h_t^{\text{none}} = 0$ and renormalize the shares to sum to 1. These instances are rare and small in absolute value.

[15] See Appendix B for more details on the calculation of average years of schooling.

[16] However, we do use the information of the birth state at time t. If selective migration by education is important, then states that have higher shares of the more mobile education category will be disproportionately represented as birth states. We assume, and the later analysis supports the idea that secondary exposed and higher education exposed workers appear to be more mobile than those with only primary or no education.

[17] In 2000, data availability is limited. The census reports the fraction of a state's residents that were born in that state, S$_{ii}$, and the fraction that is foreign born S$_{i,for}$. However, for those residents of a state who were not born in that state (S$_{ij}$, $j \neq i$, $j \neq for$), only the census region of birth is given. Conditioned on living in state i and being born in census region k, we assume the probability of having been born in state j is equal the population of state j divided by the population of region k. We make the necessary adjustment when the region of birth contains the state of residence. As data is not available for 1840, we assume the shares in 1840 are identical to the values in 1850. Also, data is not available for Alaska and Hawaii in 1940, and 1950. We assume these shares are identical to the values in 1960. For non-Census years, we linearly interpolate the shares born in state j residing in state in in year t.

schooling. For years prior to 1940 we assume that foreign born workers have the average $\overline{k}$th percentile, where the average is for the 1940–2000 period, and is state specific.[18]

To illustrate our years of schooling measure, Fig. 1 displays the average years of schooling in the labor force by census region.[19,20,21] While initial conditions certainly come into play in the first few years, within 20 years, the initial conditions have little impact. Thus New England, the Middle Atlantic and Pacific regions were clearly education leaders in the US. Except for the Middle Atlantic in 1940, all three regions remain above the average years of schooling in the US throughout the entire 1840 to 2000 period. Figure 1 indicates that the East North Central and, by 1880, the West North Central were educational leaders as well. From 1880 to 2000 the labor forces of these five regions were better educated than the average person in the labor force in the US. In contrast, the South Atlantic, East South Central and West South Central regions were educational laggards. They start with less schooling than the average in the US and remain below average throughout the data. However by 2000, these three regions have closed the gap between themselves and the US. Figure 1 also illustrates the different behavior of the Mountain region. Unlike the Pacific region which remained above the US average, the Mountain region initially lagged behind the US, and in fact lagged behind the southern states from roughly 1850 to 1870. However from 1940 to the present the Mountain region was either at or above the US average in schooling. These results are summarized in Table 1.

Table 1 contains the labor force weighted average years of schooling for each of the nine census regions and the average for the US for various years. For the US as a whole, the typical worker in 1940 had completed primary schooling and almost half a year of high school. By 1980 the typical worker was nearly a high school graduate. In 2000 the labor forces in all regions have average schooling above 12 years.

The scarcity of state-level educational estimates is our motivator, and at the same time limits our ability to verify our estimates at the state level.[22] We can, however, check the validity of our educational measures by comparing our results with previous studies estimating the average years of schooling at the national level. Fishlow (1966) used Census data before 1940 to calculate the national stock of education for both 1860 and 1900. For 1860, Fishlow estimated years of schooling for the nation of 2.06, just .02 years greater than our estimate of 2.04. For 1900, he determined the national average years of schooling was 4.96, just .02 years greater than our estimate of 4.94. Schultz (1961), following the earlier work of Long (1958), used information in the 1940 Census (the first to report years of schooling) on schooling by age cohort to backward project the national stock of education for previous census years back to 1900. For 1900 Schultz estimated that the average years of schooling was 4.14 years.[23] Our national estimate in 1900 of 4.94 is about 19%, or .8 years, greater than reported by Schultz. Therefore, our national estimate for 1900 lies between the estimates of Schultz and Fishlow.

[18] Details are in Appendix B. For information on how well our measure matches the Census data from 1940 to 2000 see Appendix D.

[19] Rather than presenting graphs with 50 lines or tables with 50 rows, aggregation at the census region is a parsimonious manner to present the data. For empirical sections, we use state level data.

[20] For a listing of states within each region, see Appendix A.

[21] We do present information about maximum gaps between states in some of our tables.

[22] Appendix D presents the comparison and contrast of our state education estimates, years of schooling, share exposed to primary and no more, share exposed to secondary and no more, and share exposed to higher education, with those of the census for 1940–2000. We feel that our estimates stand up well with the census data.

[23] Schultz (1961) reports these results in Table 7 on page 68.

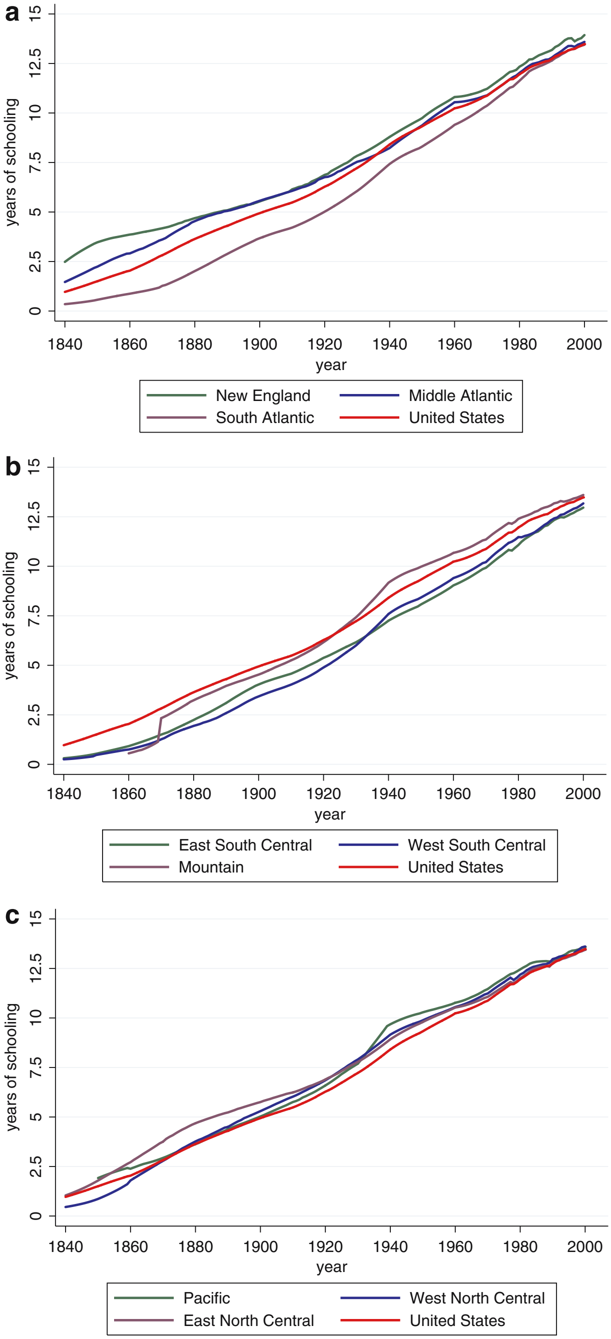

Fig. 1 Average years of schooling of the labor force by region

Table 1 Average years of schooling in the labor force (regional leaders in bold)

	1840	1860	1880	1900	1920	1940	1960	1980	2000
United States	0.97	2.04	3.64	4.94	6.28	8.41	10.2	12.0	13.5
New England	**2.48**	**3.86**	**4.69**	5.53	6.88	8.79	**10.8**	12.3	**13.9**
Middle Atlantic	1.46	2.91	4.54	5.57	6.76	8.23	10.5	12.0	13.6
South Atlantic	0.35	0.87	2.02	3.68	5.02	7.43	9.40	11.6	13.4
E. South Central	0.31	0.92	2.24	4.03	5.38	7.25	9.03	11.1	13.0
W. South Central	0.25	0.75	1.94	3.43	4.89	7.59	9.40	11.5	13.2
Mountain	–	0.55	3.23	4.53	6.17	9.17	10.7	12.4	13.6
Pacific	–	2.39	3.63	5.03	6.59	**9.68**	10.8	**12.5**	13.6
W. North Central	0.46	1.80	3.77	5.30	6.85	9.16	10.6	12.2	13.4
E. North Central	1.04	2.72	4.69	**5.75**	**6.89**	8.92	10.5	12.0	13.5
max. region gap	2.23	3.31	2.75	2.32	2.00	2.43	1.77	1.38	0.97
state max.	2.98	4.64	5.38	6.15	7.32	10.7	11.6	13.0	14.6
state min.	0.22	0.20	0.93	2.60	3.79	6.16	8.65	10.3	11.8

Table 2 Maximum schooling gaps between regions and states

	1880	1890	1900	1910	1920	1930	1940	1950	1960	1970	1980	1990	2000
R	2.75	2.64	2.32	2.21	2.00	1.90	2.43	2.21	1.77	1.54	1.38	0.97	0.97
S	4.44	4.05	3.55	3.60	3.53	3.85	4.55	3.92	2.92	2.76	2.64	2.30	2.79

Table 2 presents the maximum gap between regions, in the row marked R, and states, in the row market S, at the decadal frequency, since 1880. Table 2 illustrates the long run convergence across states and regions.[24] In 1880 the maximum gap between regions, 2.75 years, existed between the New England and West South Central regions. We pick 1880 as this is likely to be the first year in which initial conditions have no effect on the estimates. By 1900 the maximum gap between regions dropped to 2.32 years and existed between the East North Central and West South Central regions. From 1900 to 2000 the educational gap continues to narrow, reaching a nadir of 0.97 years in 2000.

The differences in average years of schooling between regions are the result of systematic differences in enrollment rates across regions. The New England, Middle Atlantic, Pacific, East North Central and, with a short lag, West North Central regions led the nation in educational attainment. These regions were the first to provide universal primary schooling, universal secondary schooling, and near universal higher education. In contrast, the South Atlantic, East South Central and West South Central regions lagged behind the country in each of these education categories. Finally the Mountain region is in between these two extreme groups.

Figure 2 illustrates the average fraction of the labor force that has been exposed to primary school, but no more. From 1840 until about 1920 the South Atlantic, East South Central and West South Central regions display shares of the labor force with elementary schooling expo-

[24] This is consistent with the convergence in enrollment rates, days attended, class size and relative teacher salaries across states from 1880–1990 in Tamura (2001).

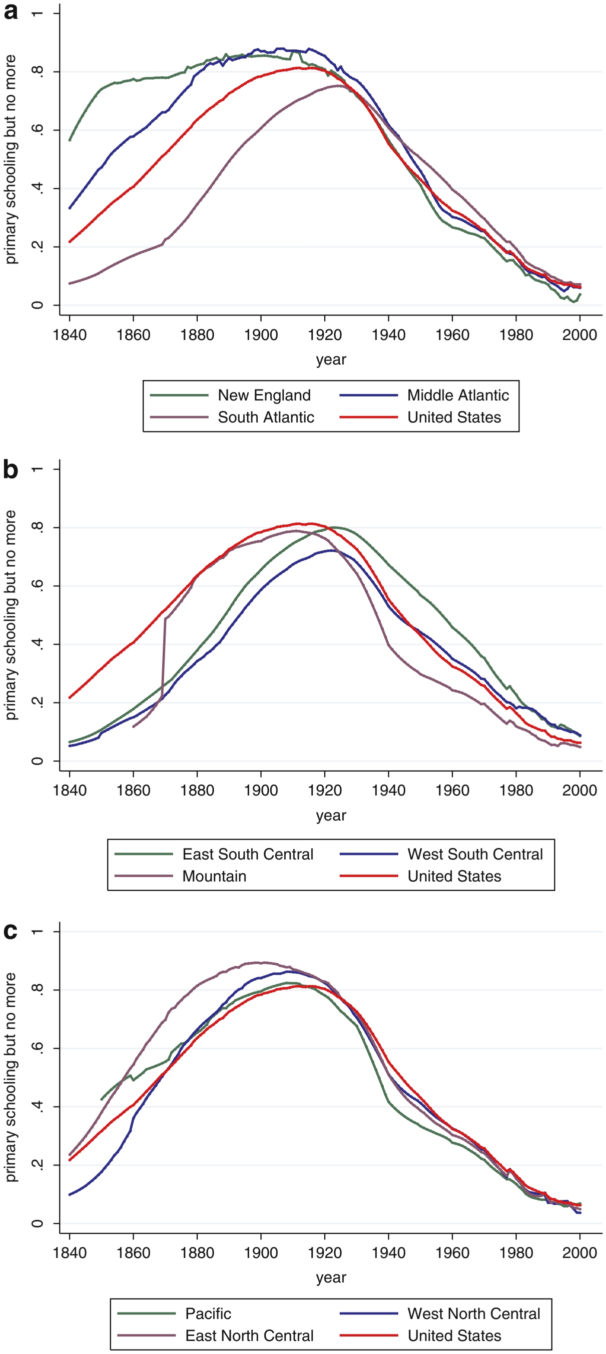

Fig. 2 Fraction of the labor force exposed to primary schooling, but no more

sure that are lower than the national average. All three are higher than the national average after 1950.[25] The New England, Middle Atlantic, East North Central and to a slightly lesser degree the West North Central have a higher share of the labor force with elementary schooling exposure than the national average from 1840 (roughly 1870 for the West North Central) until the early part of the 20th century.

Figure 3 illustrates the evidence of secondary schooling exposure but no more. For secondary schooling exposure and no more, the nine census regions behave much like they do in elementary schooling exposure. From 1840 to 1960, the New England and East North Central regions, and from 1900 to 1950 the Pacific and West North Central regions display higher than average shares exposed to secondary schooling. As Goldin (1999) and Goldin and Katz (2000) have shown, these regions were the leaders of the high school movement in the US as well as the world. The South Atlantic, East South Central, West South Central regions all lagged behind the average for the US from 1840 to the present.

The graphs displayed in Fig. 4 present the evidence for higher education. The regions with higher shares of the labor force exposed to higher education are New England, West North Central, Mountain and Pacific. The Middle Atlantic, East South Central and West South Central regions remain below average throughout the entire time period. The South Atlantic and East North Central regions seem to almost mimic the national average.

3 State per worker output

This section presents both original and existing data on state per worker output converted into real 2000 dollars.[26] In addition to the work of Easterlin (1960a,b), who provides per capita income in 1840, 1880, 1900, and 1919–1921 (1920), and government data from 1929 to 2000, we add our original estimates of real state per worker output for 1850, 1860, 1870, 1890, and 1910. Our work uses government sources to produce estimates of real agricultural output, manufacturing output, and mining output for each state for these years.[27] In combination with our measures of the labor force and the sectoral allocation of the labor force, we construct estimates of the non-agricultural, non-manufacturing non-mining output.[28] With these estimates we create output per worker by state. The details of these calculations are in Appendix C. We note that the data from 1840 to 1920 are state output per worker, while from 1929 to 2000, the data are state income per worker.

[25] In early periods, regions with large shares of the labor force exposed to elementary education are educational leaders. However, as these states are the first to have a significant fraction of their labor force exposed to secondary education, having a *smaller* fraction of the labor force exposed to elementary school later in the period is evidence of educational leadership.

[26] We convert all nominal values into real 2000 dollars, using the GDP deflator data from Gordon (1999) for years 1870–2000. For values between 1840 and 1869 we use the wholesale price index from the *Historical Statistics of the United States: Colonial Times to 1970* to compute inflation rates over this period. We then use the calculated wholesale price inflation to create a GDP deflator for the 1840–1869 period. To account for regional price differences, we use Berry et al. (2000), Mitchener and McLean (1997), and Williamson and Linder (1980). The first deflators provide measures of output or income in constant national dollars and the regional price corrections adjust for regional price variation. For the 1840–1880 period we extrapolated the trend in relative price levels for the Mountain and Pacific region. Thus the output measures are best thought of as real income per worker. More details on price level are available in Appendix B.

[27] We thank an anonymous reviewer for pointing out Towne and Rasmussen's (1960) work on agricultural value added.

[28] We thank an anonymous reviewer for referring us to Weiss (1999), which addresses methodological concerns with early US Census estimates and provides improved labor force estimates.

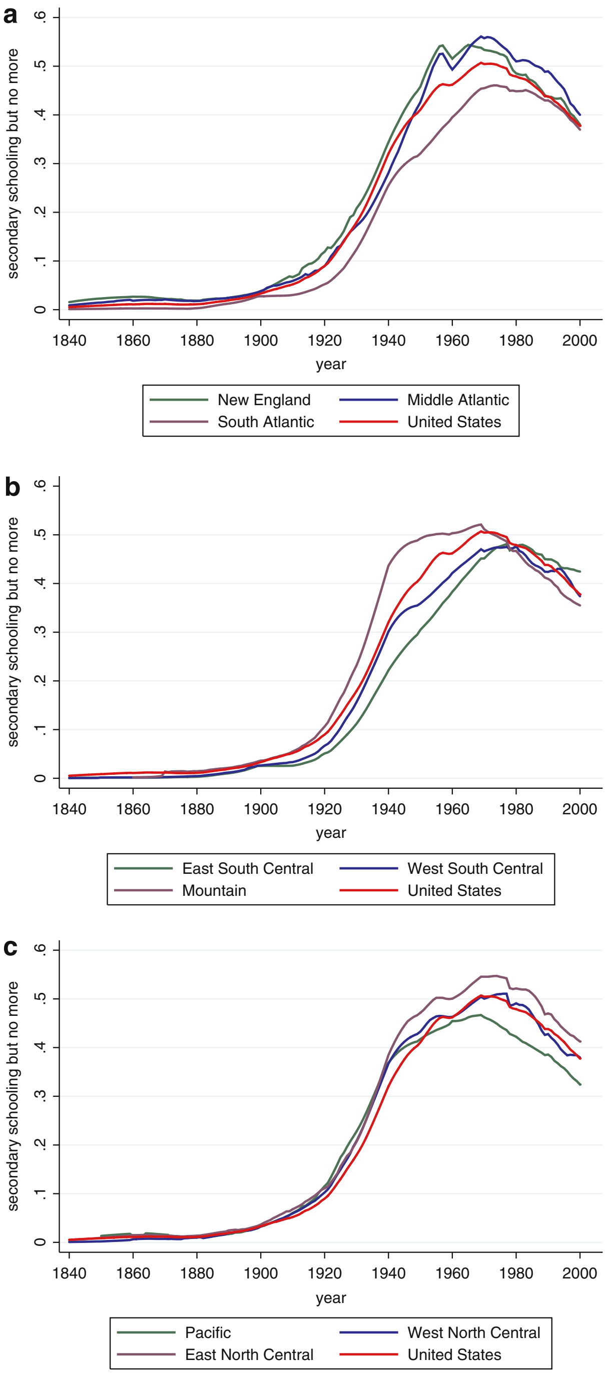

Fig. 3 Fraction of the labor force exposed to secondary schooling, but no more

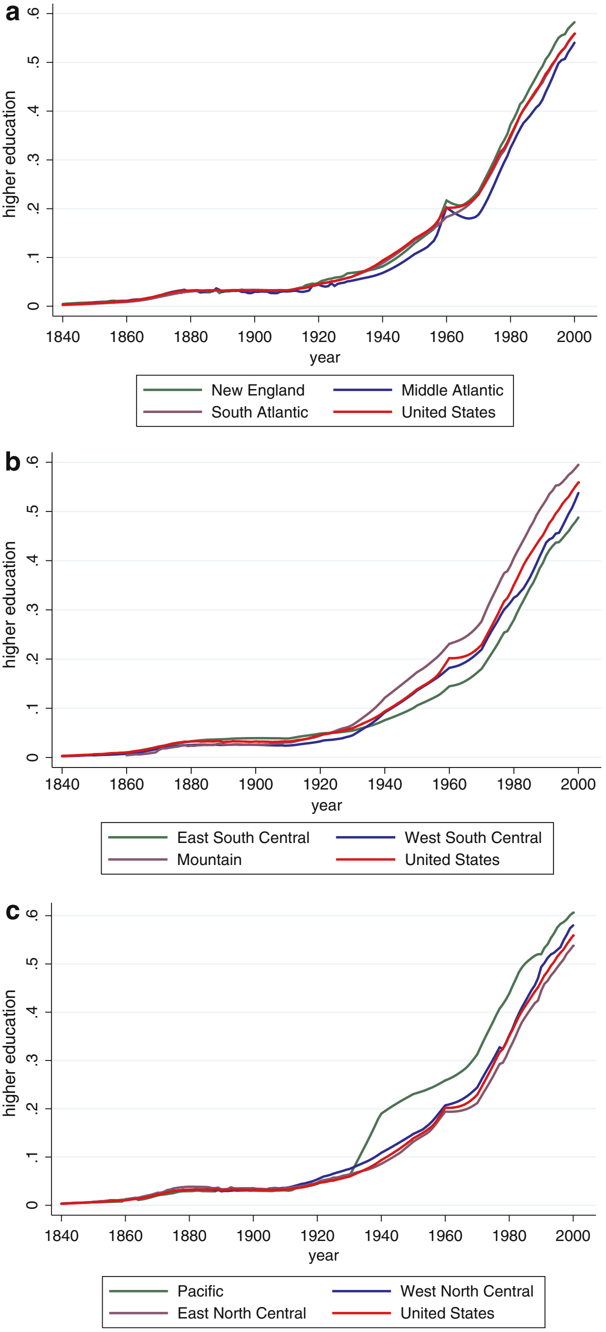

Fig. 4 Fraction of the labor force exposed to higher education

Figure 5 displays the average output per worker in each census region and the national average output per worker. As with the educational measures, we present the data in regional aggregates in order to easily facilitate data presentation. The real income per worker series has many similarities with the educational attainment data. The Middle Atlantic and Pacific regions are consistently more productive than the US from 1840 to 2000, and the South Atlantic, East South Central, and West South Central regions are consistently less productive than the US from 1840 to 2000. The remaining three regions, Mountain, West North Central, and East North Central are essentially as productive as the US from 1840 to 2000.

As apparent from Fig. 5 as well as Table 3, real output per worker has increased substantially in the US, and within all regions. Consistent with evidence for the US from Baier et al. (2006), real output per worker grew at an annual rate of 1.66% per year. The nine census regions had annual real output per worker growth rates of 1.54 (New England), 1.54 (Middle Atlantic), 2.03 (South Atlantic), 1.68 (East South Central), 1.55 (West South Central), 1.14 (Mountain), 0.66 (Pacific), 1.68 (West North Central) and 1.55 (East North Central).[29]

The low growth rate values for the Mountain and Pacific regions are due to unique factors. For the Mountain region in 1850, only New Mexico and Utah are in the data. Each has worker productivity in excess of 10,000 dollars, well above the US value of 6,691 dollars. By 1870 all states are included in the regional calculation. The additional information on the productive mining states of Colorado and Nevada, with worker productivity in excess of 20,000 dollars, generates a relatively high initial value for income per worker.

The 1860 Pacific region calculation consisted of California, Oregon and Washington. All three reported real output per worker values in excess of 15,000 dollars. These states were likely very high cost of living states as many manufactured goods would have to be imported from the rest of the US or abroad. Real output per worker for the Pacific region grows at an annual rates of 1.24 percent from 1880 to 2000, 1.41% from 1900 to 2000, and 1.56% per year from 1920 to 2000.

Our results are consistent with Goldin and Margo (1992) over the 1840–1860 period. They report falling real wages for artisans in three of four regions, and stagnant or falling real wages for laborers and clerks in three of four regions.[30] We find that for the typical US non-agricultural worker, output per worker is roughly constant (including California), and slightly falling at an annual rate of .4% (excluding California). Goldin and Margo find that real wages for artisans and laborers clearly rise in the Northeast, with stagnant real wages for clerks. Equally weighting their three groups produces annualized real wage growth of about .8% per year. We find rising real output for non-agricultural workers in the Northeast at .3% per year. We find falling real output per non-agricultural worker in the South Central states of roughly 1.6% per year, similar to the 1% decline in real wages of artisans, clerks and laborers found by Goldin and Margo for the same region. Goldin and Margo identify essentially constant real wages for artisans, clerks and laborers in the South Atlantic region, again consistent with our results.[31] Finally in the Midwest region, Goldin and Margo show falling real wages for artisans and laborers, roughly 1% per year, and rising real wages for

[29] The Mountain region is from 1850 to 2000, and the Pacific region is from 1860.

[30] Goldin and Margo report artisan, laborer and clerk wages in the Northeast, Midwest, South Atlantic and South Central in Tables 2A.5–2A.7. To match these geographic regions we combined New England and the Middle Atlantic states to create the Northeast; we combined the East North Central and West North Central states to create the Midwest, and combined East South Central and West South Central to produce the South Central region.

[31] For the South Atlantic region, averaging across states produces a .2% annual decline in real output per worker, weighting by non-agricultural workers produces a .4% annual increase in real output per worker.

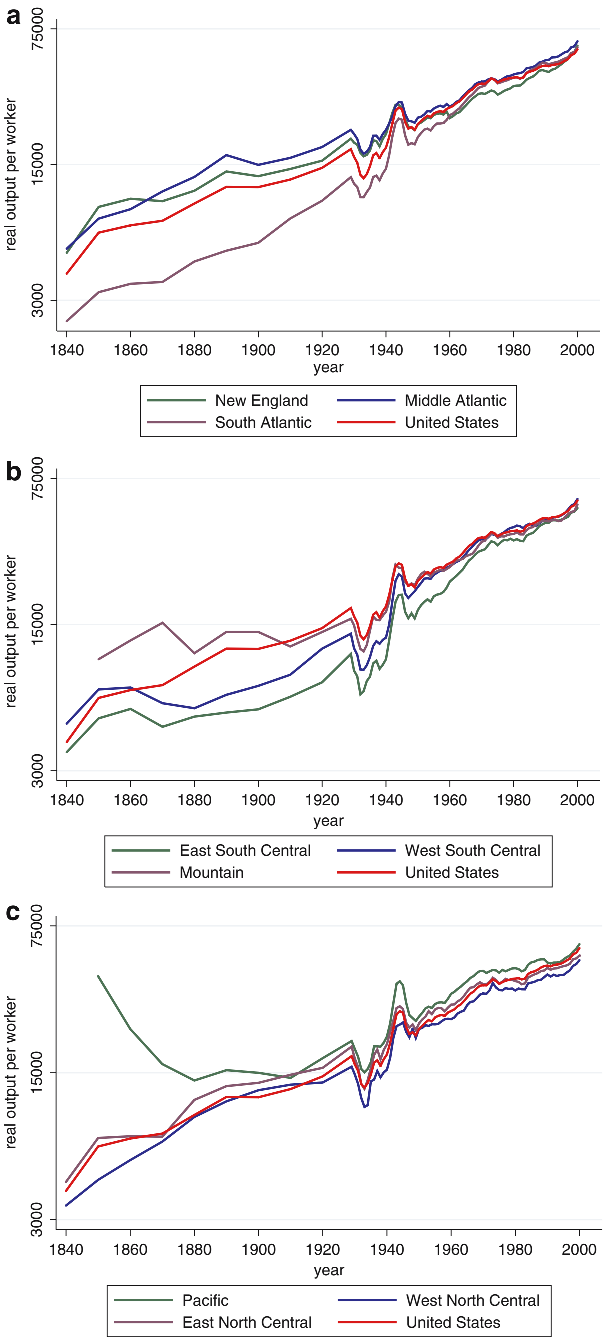

Fig. 5 Real output per worker, by region

Table 3 Real output per worker (regional leaders in bold)

	1840	1860	1880	1900	1920	1940	1960	1980	2000
United States	4,114	7,297	9,449	11,477	14,429	18,328	29,514	42,083	58,791
New England	5,267	9,999	10,998	13,073	15,706	21,518	26,042	38,074	61,426
Middle Atlantic	**5,528**	8,840	12,954	14,947	**18,469**	**22,639**	29,854	43,667	**64,758**
South Atlantic	2,342	3,647	4,752	5,929	9,770	14,278	26,982	42,058	60,216
E. S. Central	3,683	5,928	5,447	5,900	7,947	10,240	24,092	37,899	54,134
W. S. Central	5,042	7,503	5,971	7,641	11,512	12,993	28,521	43,845	59,833
Mountain	–	12,606	10,951	13,838	13,823	17,247	28,272	40,690	56,277
Pacific	–	**24,257**	**13,786**	**14,992**	17,606	22,302	**35,638**	**47,185**	61,374
W. N. Central	3,503	5,760	9,248	12,395	13,486	15,515	26,991	36,952	51,527
E. N. Central	4,540	7,484	11,147	13,440	15,842	20,512	31,641	40,972	54,162
Region $\frac{\max}{\min}$	2.36	6.65	2.90	2.54	2.32	2.21	1.48	1.25	1.26
State max.	6,820	25,185	18,991	17,088	20,492	28,797	38,531	62,117	82,438
State min.	1,990	2,984	3,297	3,678	6,019	7,135	20,032	31,558	41,653
State $\frac{\max}{\min}$	3.43	8.44	5.76	4.65	3.40	4.04	1.92	1.97	1.98

clerks, 1.8% per year. We find larger annualized declines in real output per non-agricultural workers of 3%, weighting by non-agricultural workers and .6% unweighted state average. In contrast we find that real output per agricultural worker doubles between 1840 and 1860, although their share of the labor force declines from 80% to 56%.

The effects of the Civil War are quite prominent in the figures, and are evident in Table 3. The states of the old Confederacy, South Atlantic, East South Central and West South Central clearly have lower growth rates. Between 1860 and 1880, these three regions experienced real annual income per worker growth of 1.32%, −0.42% and −1.14%, respectively. The annual growth rates of income per worker from 1860 to 1870 for these three regions are 0.22%, −1.97% and −1.73%, respectively. In 1860 their relative worker productivity values are 50%, 81%, and 103% of the national average, while in 1880 their relative productivity fell to 50%, 58%, and 63%, respectively. By 2000 only the East South Central remains below the national average.

The final four rows of Table 3 present evidence on regional output per worker convergence. These contain the ratio of the maximum regional income per worker to minimum regional income per worker, the maximum and minimum state per worker income, and the ratio of the maximum state income per worker to minimum state income per worker. Inequality in 1870 and 1880 are certainly higher than in the pre Civil War period. Inequality in output per worker is reduced throughout the next century. By 1980 the relative regional gap and the relative state gap is about one third of its value in 1880. However, both the relative state gap and regional gap have increased somewhat from 1980 to 2000.[32]

4 Robustness check: returns to schooling

Though our estimated years of schooling appear similar to national estimates by Schultz and Fishlow, we also estimate returns to state-level measures of schooling to determine if our measures exhibit reasonable returns. Before we present evidence on the rate of return to schooling, it is necessary to deal with missing data on other inputs. Consider a model with two factors of production, human capital and all other inputs which we call physical capital. We assume production of a single final output is Cobb–Douglas. We assume perfect competition in factor markets and free mobility of capital. Output per worker in state i is given by

$$y_{it} = A_{it}k_{it}^{\alpha}\,(human\ capital)_{it}^{1-\alpha} \tag{17}$$

where k_{it} is physical capital per worker and $human\ capital_{it}$ is human capital per worker. Under perfect competition in the output market, with final output as numeraire, the representative firm solves:

$$\max_{k_{it},h_{it}} \left\{A_{it}k_{it}^{\alpha}\,(human\ capital)_{it}^{1-\alpha} - r_t k_{it} - w_t human\ capital_{it}\right\} \tag{18}$$

where r_t and w_t are the rental rate per unit of physical capital and human capital, respectively. Under competition firms choose physical capital in proportion to the human capital in the workforce:

$$k_{it} = \left(\frac{w_t}{r_t}\right)\left(\frac{\alpha}{1-\alpha}\right) human\ capital_{it} \tag{19}$$

[32] These results are consistent with those found using state income per capita from 1880, 1900, 1920 and 1930–1990 at the decadal frequency in Tamura (2001).

Therefore substituting this back into the output equation produces:

$$y_{it} = A_{it}\left(\frac{w_t}{r_t}\left(\frac{\alpha}{1-\alpha}\right)\right)^{\alpha} human\ capital_{it} \tag{20}$$

We assume that $human\ capital_{it}$ can be specified in a Mincerian fashion:

$$human\ capital_{it} = \exp\left(\beta E_{it} + \gamma x_{it}\right) \tag{21}$$

where E_{it} is years of schooling in state i in year t, and x_{it} is experience in state i in year t.[33] In order to construct average experience by state, we calculated average age in the state per persons not enrolled in school and under the age of 65. From average age we subtract the sum of our average years of schooling measure in the labor force and the 6 years before individuals typically begin school enrollment. With this definition of $human\ capital_{it}$ the "earnings regression" is:

$$\ln y_{it} = \ln A_{it} + \alpha \ln\left(\frac{w_t}{r_t}\left(\frac{\alpha}{1-\alpha}\right)\right) + \beta E_{it} + \gamma x_{it} \tag{22}$$

Identification of β in (22) requires assumptions on the nature of state specific levels of Total Factor Productivity, A_{it}, as well as the national wage-rental ratio. If we assume that each state has a common level of TFP, and that labor and physical capital are perfectly mobile, then we can estimate (22) using time dummies in a pooled time series cross section. The coefficient on years of schooling identifies β.[34]

Table 4 contains the results of real per worker output regressed on years of schooling.[35] The first four columns include year dummies to allow for more variation in technological change than a deterministic trend. The second column allows for a different intercept for the

[33] Those familiar with the standard Mincer earnings regression may wonder why we exclude the quadratic term in experience. This is because of aggregation bias. While one can construct a model in which the linear terms in education and experience are identified by state variation, the quadratic term is not identified upon aggregation. When we experimented with identification, the results confirmed the bias in estimation, and hence we ignore the diminishing returns to work experience. The results indicate that experience returns are significantly below that from additional schooling and hence suggest that ignoring the quadratic term is not problematic.

[34] If we drop the assumption that labor and physical capital are perfectly mobile across state boundaries, (22) becomes:

$$\begin{aligned}\ln y_{it} &= \ln A_{it} + \alpha \ln\left(\frac{k_{it}}{human\ \ capital_{it}}\left(\frac{\alpha}{1-\alpha}\right)\right) + \beta E_{it} + \gamma x_{it} \\ &= \ln A_{it} + \alpha \ln k_{it} + \alpha \ln\left(\frac{\alpha}{1-\alpha}\right) + \beta(1-\alpha)E_{it} + \gamma(1-\alpha)x_{it}\end{aligned}$$

We can estimate the above equation with state specific time trends and time dummies. Our estimate on years of schooling will be a combination of both the rate of return to schooling and labor's share of income. Therefore we need an estimate of the share of output that labor receives, $(1-\alpha)$. Table B2 in Appendix B presents evidence on labor's share which varies between $\frac{2}{3}$ and $\frac{4}{5}$ throughout the 100 years of observations. Our estimates using this methodology are generally higher than those presented in the paper. We do not report them for brevity, they are available on request from the authors.

[35] The coefficient estimates in Tables 4 and 5 are the results of weighted least squares regressions where we use the labor force as the weight. This is appropriate if, at the individual level, the data satisfies the homoskedasticity assumption. In this case, the variance of the error term will be $var(\varepsilon) = \sigma^2/L^2$. Weighting by the labor force corrects for the heteroskedasticity of known form. We also ran the weighted least squares regressions where we computed robust standard error. All coefficient estimates remained statistically significant. In addition, standard ordinary least sqaures regressions produced results that were not qualitatively different from the weighted least squares regressions. These sensitivity results are available upon request.

Table 4 Earnings regressions: annual data (standard error)

E	.1517	.1506	.1506	.1506	.1517	.1506	.1506	.1506
	(.0035)	(.0035)	(.0035)	(.0035)	(.0034)	(.0034)	(.0034)	(.0034)
Exp.	.0324	.0341	.0343	.0344	.0324	.0341	.0343	.0344
	(.0017)	(.0018)	(.0018)	(.0018)	(.0017)	(.0018)	(.0017)	(.0017)
N	4004	4004	4004	4004	4004	4004	4004	4004
$\overline{R}^2$	.9232	.9239	.9240	.9241	.4040	.4097	.4105	.4107
Year dummies	Yes	Yes	Yes	Yes	No	No	No	No
Ak intercept	No	Yes	Yes	Yes	No	Yes	Yes	Yes
Ak E	No	No	Yes	Yes	No	No	Yes	Yes
Ak exp.	No	No	No	Yes	No	No	No	Yes
Differenced	No	No	No	No	Yes	Yes	Yes	Yes

Alaska. The third column allows for a different intercept and different return to schooling for Alaska. The fourth column allows for a different intercept for Alaska, as well as different returns for both schooling and experience for Alaska. Under the hypothesis that TFP does not differ across states, i.e., $A_{it} = A_t$ for all i, differencing each state's log output per worker from the labor force weighted log US output per worker, years of schooling, and average experience from the labor force weighted US averages allows for the estimation of (Eq. 22) without any time controls. These differenced regressions are reported in the final four columns of Table 4.

Given the specification implied by Eq. 23, the results in Table 4 indicate an overall return to schooling, including the implied physical capital return, of 15% per year of schooling. These results are consistent with the evidence presented in Angrist and Krueger (1991), Staiger and Stock (1997), and Card (1995). The returns to experience, reflecting on-the-job training or learning by doing, are similar across all four columns. A one year increase in average experience raises worker productivity by about 3%.

Failing to account for the rising female labor force participation rate present over this period may result in poor estimates. To control for this we correct for the share of the labor force that is female (male) and interact these shares with average years of experience. This allowed us to separately measure the rate of return to experience for each sex. The results of these are contained in Table 5. The second through fifth column reports the average return to schooling and average estimated returns to experience by sex with varying controls for Alaska. The last four columns are the differenced regressions, as in Table 4. The rows marked F and Prob > F contain the F statistic on the test of equality of returns to experience between men and women, and the p-value of the statistic.

The results of Table 5 indicate that the estimated returns to schooling are robust to the possible differences in returns to experience between men and women. It is reasonable to state that an additional year of schooling in a randomly chosen state returns 15%. Rates of returns to experience for men and women are statistically different in all seven regressions. The typical male worker becomes about 4% more productive at the individual level per additional year of experience, whereas the typical female worker only gains 3% in productivity per additional year of experience.

One might be concerned that our estimates of the return to schooling may be biased because we assume a common intercept for all states in any time period. To address this

Table 5 Earnings regressions: annual data (standard error)

E	.1500	.1489	.1489	.1488	.1500	.1489	.1489	.1488
	(.0036)	(.0036)	(.0036)	(.0036)	(.0035)	(.0035)	(.0035)	(.0035)
Exp male	.0375	.0394	.0396	.0397	.0376	.0395	.0397	.0398
	(.0030)	(.0030)	(.0030)	(.0030)	(.0030)	(.0030)	(.0030)	(.0030)
Exp female	.0270	.0285	.0287	.0288	.0267	.0282	.0283	.0284
	(.0032)	(.0032)	(.0032)	(.0032)	(.0031)	(.0031)	(.0031)	(.0031)
N	4004	4004	4004	4004	4004	4004	4004	4004
$\overline{R}^2$	.9233	.9240	.9241	.9241	.4043	.4101	.4108	.4111
F	4.17	4.57	4.57	4.62	4.59	5.08	5.07	5.13
Prob($>F$)	.0412	.0325	.0326	.0316	.0322	.0243	.0244	.0236
Year dummies	Yes	Yes	Yes	Yes	No	No	No	No
Ak intercept	No	Yes	Yes	Yes	No	Yes	Yes	Yes
Ak E	No	No	Yes	Yes	No	No	Yes	Yes
Ak exp.	No	No	No	Yes	No	No	No	Yes
Differenced	No	No	No	No	Yes	Yes	Yes	Yes

concern, one way to correct for this is to allow for state specific effects. To help guide our thinking about alternative specifications that would correct for this potential bias, we return to Eq. 22

$$\ln y_{it} = \ln A_{it} + \alpha \ln\left(\frac{w_t}{r_t}\left(\frac{\alpha}{1-\alpha}\right)\right) + \beta E_{it} + \gamma \exp_{it} \tag{23}$$

One way to rewrite the above specification in a form that allows for different intercepts is to assume that the distribution of state specific technology is constant over time and there are systemic time effects. The regression specification implied from Eq. 22 is, therefore, given by

$$\ln y_{it} = c_i + b_t + \beta E_{it} + \gamma \exp_{it} + u_{it} \tag{24}$$

where c_i is the state specific fixed effects and b_t is a time specific effect common to all states. One way to interpret the above equation, in the context of Eq. 22, is that the state specific technology is given by $\ln(A_{it}) = c_i + u_{it}$ and that there are national labor and capital markets so that $\frac{w_t}{r_t} = b_t$. To correct for the state specific effects, there are two standard approaches to adjust for these effects: fixed effects regressions or OLS on first-differenced data. In both cases, it is required that there are no feedback effects from innovations in income to future levels of educational attainment. If this is the case, then standard fixed effects regressions or first-differencing leads to inconsistent estimates of the return to schooling.

The first column of Table 6 reports the results of a standard fixed effects regression on the decadal years from 1860 to 2000.[36] With fixed effects we find the return to education is roughly 12% per year and that the return to experience is not significantly different from zero. Since it is likely that there is autocorrelation in the data, column (2) presents fixed effects estimation with autocorrelated errors. However, since educational attainments decisions may respond to expected changes in income and because income growth may lead to

[36] We only used data for census years due to the high degree of serial correlation.

Table 6 Fixed effects with leads of education (standard error)

	(1)	(2)	(3)	(4)
E_t	0.120	0.100	−0.039	0.050
	(0.018)	(0.025)	(0.039)	(0.030)
E_{t+1}			0.156	0.095
			(0.039)	(0.031)
Exp	−0.035	0.088	−0.044	−0.215
	(0.076)	(0.018)	(0.085)	(0.059)
N	718	667	667	616
Decade dum.	Yes	Yes	Yes	Yes
AR errors	No	Yes	No	Yes

more educational attainment, we need to be concerned about fixed effects and the presence of feedback effects (see Wooldridge, 2002). To test for possible feedback effects, we follow Wooldridge (2002) and run a fixed effects regression with a lead of educational attainment in the specification. If the coefficient on future educational attainment is statistically different from zero, then we will take this as evidence that contemporaneous innovations in income lead to future educational attainment. These results are reported in column 3 and 4 of Table 6. In column 3 the results are for the specification with a forward lead and *without* the autoregressive component and column (4) presents the results with the lead with the autoregressive errors.

With time dummies, the return to schooling from the fixed effects regression is roughly 12%. When we add a lead of education to the fixed effects regression, the return to contemporaneous education is negative and insignificant and the return to the lead of education is positive and significant at the 5% level. In the autoregresive specification, the return to contemporaneous education is significant at the 10% level and the lead of education is significant at the 5% level. Therefore, at the 5% level we cannot reject the null hypothesis of no feedback effects. Since the p-value is sufficiently low, we would like allow for the possibility of feedback effects from current income to future education and experience.

If feedback effects are present, the standard approaches to correct for state effects will lead to inconsistent estimates. To correct for the possibility of state specific effects and the autoregressive nature of the error term, we follow Blundell and Bond (1998, 1999) and rewrite Eq. 22 as

$$\ln y_{it} = c_i + b_t + \beta E_{it} + \gamma x_{it} + u_{it} \tag{25}$$

$$u_{it} = \rho u_{it-1} + e_{it} \tag{26}$$

The above expression can be rewritten as:

$$\begin{aligned}\ln y_{it} = &(1-\rho)\, c_i + b_t - \rho b_{t-1} + \rho \ln y_{it-1} + \beta E_{it} - \rho\beta E_{it-1} + \gamma x_{it} \\ &-\rho\gamma x_{it-1} + e_{it}\end{aligned}$$

Table 7 System GMM dynamic panel estimation (standard error)

	IV educ	IV educ	IV educ	IV educ	IV educ	IV educ
E_{it}	0.117	0.117	0.114	0.127	0.128	0.127
	(0.028)	(0.027)	(0.027)	(0.028)	(0.028)	(0.028)
E_{it-1}	0.002	−0.001	−0.002	0.002	0.001	−0.001
	(0.030)	(0.029)	(0.030)	(0.030)	(0.030)	(0.030)
$\ln y_{it-1}$	0.571	0.566	0.568	0.564	0.559	0.556
	(0.067)	(0.066)	(0.064)	(0.071)	(0.069)	(0.068)
N	667	667	667	667	667	667
Instruments	2 lags	3 lags	4 lags	2 lags ed, exp,diff-ed	3 lags ed, exp,diff-ed	4 lags ed, exp,diff-ed
Decade dum.	Yes	Yes	Yes	Yes	Yes	Yes

so the estimating equation becomes:

$$\ln y_{it} = (1-\rho)\,c_i + b_t - \rho b_{t-1} + \pi_1 \ln y_{it-1} + \pi_2 E_{it} + \pi_3 E_{it-1} + \pi_4 x_{it} \\ + \pi_5 x_{it-1} + e_{it}$$

As in Bond and Blundell (1998), we use differenced and lagged values of the data as instruments in the levels regression. As additional instruments, we experimented with lags of the difference between state i's average educational attainment and the average educational attainment of the other states in the region—this variable may capture the changes in educational attainment related to regional convergence. More specifically, we create the variable:

$$E_{it}^{c} = \left[E_{it} - \frac{1}{N^R - 1} \sum_{j \neq i}^{N^R} E_{jt} \right] \tag{27}$$

In all the regressions, average experience was determined to be collinear with the other right-hand side variables and it was subsequently dropped from the specification.

In the above specifications, the return to education ranges between 11% and 13% all within the range of most microeconometric estimates. In none of the specifications, was the lag of educational attainment significant. Given the structure of the model, the coefficient on lagged education (π_3) should equal $-\pi_2\pi_1$. Thus, given the other coefficient estimates this implies that the empirical estimate of π_3 should equal (roughly) −0.065. In the above specifications, we cannot reject, at the 5% level, that the coefficient estimates $\pi_3 = -\pi_2\pi_1$. For robustness, we employed additional lag lengths, and lag structures and all results were qualitatively similar. Thus, allowing for and correcting for feedback effects from income to education does not alter the fundamental finding that these calculated average years of schooling and income measures deliver estimates of the return to education that fall within the range of estimates found in the microeconometric literature.

5 Conclusion and extensions

Motivated by the scarcity of state-level data on education in the 19th and early 20th century, this paper employs historic state enrollment and population data to produce original average years of schooling measures for each state from 1840 to 2000. These measures will benefit economics, social science, education, or history researchers searching for consistent historic schooling measures for empirical studies. We show that there has been tremendous increases in schooling in the US over the 1840 to 2000 period, with average years of schooling rising from 1 year to over 13 years. In addition there has been a reduction in the variance across states. We also construct original estimates for state per worker output for the census years 1850, 1860, 1870, 1890 and 1910. Coupling our constructed data with previous work by Easterlin, Weiss, and government data, we produce state per worker income measures for 1840 through 1920 at the decadal frequency and 1929 through 2000 at the annual frequency. We then estimate aggregate Mincerian earnings regressions and discover that the return to a year of schooling for the average individual in a state ranges from 11% to 15%. This range is robust to various time periods, various estimation methods and various assumptions about the endogeneity of schooling.

This work is part of a larger research agenda seeking to construct state-level measures of aggregate inputs in order to perform a systematic analysis of cross-state income variation in the United States from 1840 to 2000. In a companion paper, we have computed real state physical capital per worker for the states of the United States over this same horizon, Turner, Tamura, Mulholland, and Baier (2006). Though many cross-country analyses have increased our knowledge of the importance of TFP and TFP growth in determining both the level differences in income as well as the growth rate of income and its variation, many economists, as listed in Temple (1999), object to the empirical work on growth. One objection is the inability to account for large heterogeneity in social, religious, and institutional characteristics. Another criticism is the small time frame over which cross-country inputs, income, and TFP are estimated. By creating and analyzing new state measures of human capital, physical capital, and income of the United States over 160 years, we intend to reduce both the possible problems associated with the social, religious, and institutional heterogeneity and the errors that can be induced by business cycles when comparing cross-sectional TFP over shorter time periods. Therefore, it is our hope, that these years of schooling measures may allow for a precise measure of technology growth, and with it, a more comprehensive explanation of why income variation occurs across developing counties such as the United States in the 1800s.

Following the cross-country work of Klenow and Rodriguez-Clare (1997) and Easterly and Levine (2001), we also envision future research assessing whether the variance in the growth rate of TFP may account for the majority of the variance in the growth rate of output. Three possible sources of regional TFP variation include: variation in educational attainment by race; variation in educational quality; and variation in sectoral allocation of labor. We intend to merge additional data on demographics, educational quality, and labor allocation by sector to determine the impact that variation within a region has on the growth of TFP.

Acknowledgements We thank the workshop participants at Clemson University, Texas A & M, the University of Kentucky, SUNY Buffalo, the University of Virginia, the University of South Carolina, and the 2004 Midwest Macroeconomic Meetings for helpful comments. We benefited from suggestions by Kevin Murphy, Casey Mulligan, and anonymous referees. All errors are the responsibility of the authors.

Appendix A

There are nine census regions in the US. The following Table provides the regional groups.

New England	Middle Atlantic	South Atlantic	E. South Central	W. South Central
Connecticut	New Jersey	Delaware	Alabama	Arkansas
Maine	New York	D.C.	Kentucky	Louisiana
Massachusetts	Pennsylvania	Florida	Mississippi	Oklahoma
New Hampshire		Georgia	Tennessee	Texas
Rhode Island		Maryland		
Vermont		North Carolina		
		South Carolina		
		Virginia		
		West Virginia		
Mountain	**Pacific**	**W. North Central**	**E. North Central**	
Arizona	Alaska	Iowa	Illinois	
Colorado	California	Kansas	Indiana	
Idaho	Hawaii	Minnesota	Michigan	
Montana	Oregon	Missouri	Ohio	
Nevada	Washington	Nebraska	Wisconsin	
New Mexico		North Dakota		
Utah		South Dakota		
Wyoming				

Tables A1–A3 contain average elementary school enrollment rates, secondary school enrollment rates, and higher education enrollment rates by census region as well as for the US as a whole from 1840–2000. We note that the elementary enrollment rates are often over 100%. In the early periods, higher elementary enrollment rates are due to two factors: older aged first-time enrollment and less social promotion. The methodology we present addresses a portion of these sources.

Table A1 Average elementary enrollment rates

	1840	1860	1880	1900	1920	1940	1960	1980	2000
US	48.1	74.7	95.9	106.8	108.4	104.4	100.8	100.6	105.1
New England	129.1	118.9	118.8	116.9	109.4	106.5	101.7	102.1	106.3
Mid Atlantic	75.4	94.0	114.7	108.2	103.4	106.6	104.4	98.8	105.0
So. Atlantic	13.7	28.1	73.4	95.4	106.6	101.7	98.1	101.7	106.7
E. So. Central	13.3	42.3	75.3	104.2	114.2	109.1	98.9	100.4	108.9
W. So. Central	8.2	24.9	46.4	82.9	104.4	101.1	97.1	102.5	108.2
Mountain	–	19.1	82.7	109.4	114.4	101.5	100.5	100.7	100.6
Pacific	–	70.4	108.6	121.0	122.1	106.6	99.9	100.3	103.3
W. N. Central	18.2	77.9	105.1	119.9	112.7	105.6	102.5	99.5	103.1
E. N. Central	45.6	111.8	113.9	113.6	107.2	102.7	102.0	100.0	104.3

Table A2 Average secondary enrollment rates

	1840	1860	1880	1900	1920	1940	1960	1980	2000
US	2.0	3.1	4.0	10.3	28.0	72.4	84.9	89.2	92.7
New England	4.7	4.2	4.2	20.5	40.4	78.4	88.6	90.5	94.7
Mid Atlantic	2.9	3.8	4.7	12.5	27.9	83.0	90.0	94.9	97.2
So. Atlantic	0.8	1.5	3.6	5.1	14.6	56.8	76.8	85.2	91.0
E. So. Central	0.8	2.0	3.5	4.7	12.1	44.2	74.7	83.8	88.9
W. So. Central	0.5	1.5	2.9	4.8	18.7	63.8	81.7	84.6	89.7
Mountain	–	0.9	5.0	10.5	40.4	76.9	86.3	87.6	88.6
Pacific	–	3.3	4.0	12.9	56.4	89.2	86.1	90.0	98.3
W. N. Central	0.8	3.2	3.9	11.7	37.2	78.2	90.0	91.6	93.7
E. N. Central	1.7	4.3	4.2	13.4	34.4	80.5	88.5	90.9	90.1

Table A3 Average higher education enrollment rates

	1840	1860	1880	1900	1920	1940	1960	1980	2000
US	0.7	1.4	0.9	1.5	6.2	8.4	22.2	40.4	57.0
New England	0.9	0.9	1.2	1.8	8.6	8.2	26.6	47.2	71.9
Mid Atlantic	0.6	0.7	0.7	1.2	5.9	7.7	22.5	40.1	59.0
So. Atlantic	0.6	1.7	1.2	1.3	4.8	6.4	16.1	35.6	53.9
E. So. Central	0.7	1.6	1.7	1.6	2.9	5.6	16.4	31.8	48.0
W. So. Central	1.7	1.7	0.6	0.9	3.7	8.3	20.2	33.9	47.7
Mountain	–	1.4	1.3	2.1	5.9	10.5	25.6	42.0	61.5
Pacific	–	1.5	1.1	2.2	10.2	12.9	29.5	54.1	59.8
W. N. Central	0.9	2.1	0.7	1.8	8.0	9.6	24.4	38.3	62.6
E. N. Central	0.7	1.7	0.5	1.4	7.4	9.2	22.5	39.3	58.0

Appendix B

In this Appendix we provide details on the calculations of years of schooling.

I. Description of Data
   A. Public enrollment
   B. Private enrollment
   C. Higher educational enrollment
   D. Population
   E. Labor force
   F. Price levels
   G. Expected years
II. Description of calculations
   A. Enrollment rates
   B. Eductional exposure fractions (primary, secondary, college)
      1. General methodology
      2. Higher education/Higher education inflow adjustment (Θ)

3. Secondary education/Secondary departure rate (δ)
4. Elementary and no education
5. Values of Θ and $1 - \delta$
6. Initial conditions

C. Educational exposure fractions for foreign born

III. Idiosyncrasies

A. DC/MD/VA

B. AK/HA

C. ND/SD/Dakota

D. OK/Indian Territory

IV. Labor's share of income

V. First year of data availability

Data description

Public enrollment data

Public enrollment, 1840–1916: Data for total (elementary and secondary) public enrollment are available from decennial census data, by state, in 1840, 1850, 1860, 1870. Total public enrollment data are available in *Statistical Abstracts of the United States* for the years 1872, 1877, 1879–1887, 1889–1891, and 1893–1916.

Data for total public enrollment for non-decennial years between 1840 and 1870 was log linearly interpolated. Data for the years 1871, 1873–1876, 1878, 1888, and 1892 was also log linearly interpolated.

We do not observe the fraction of total public enrollment that is elementary versus secondary until the year 1899. However, we do have national aggregates that make this breakdown in 1870, 1880, and 1890–1898.

Letting $pub.enroll_{it}^{\text{primary}}$ designate the public primary enrollment level in state i for time period t, and $pub.enroll_{it}^{\text{total}}$ refer to the total (primary and secondary) enrollment level, we assign:

$$pub.enroll_{it}^{\text{primary}} = pub.enroll_{it}^{\text{total}} \frac{\sum_j pub.enroll_{j,1870}^{\text{primary}}}{\sum_j pub.enroll_{j,1870}^{\text{total}}}, \quad t \leq 1870 \tag{28}$$

$$pub.enroll_{it}^{\text{primary}} = pub.enroll_{it}^{\text{total}} \frac{\sum_j pub.enroll_{j,1880}^{\text{primary}}}{\sum_j pub.enroll_{j,1880}^{\text{total}}}, \quad 1871 \leq t \leq 1880 \tag{29}$$

$$pub.enroll_{it}^{\text{primary}} = pub.enroll_{it}^{\text{total}} \frac{\sum_j pub.enroll_{j,1890}^{\text{primary}}}{\sum_j pub.enroll_{j,1890}^{\text{total}}}, \quad 1881 \leq t \leq 1890 \tag{30}$$

$$pub.enroll_{it}^{\text{primary}} = pub.enroll_{it}^{\text{total}} \frac{\sum_j pub.enroll_{jt}^{\text{primary}}}{\sum_j pub.enroll_{jt}^{\text{total}}}, \quad 1891 \leq t \leq 1898 \tag{31}$$

$$pub.enroll_{it}^{\text{secondary}} = pub.enroll_{it}^{\text{total}} - pub.enroll_{it}^{\text{primary}} \tag{32}$$

Beginning in 1899, we observe both $pub.enroll_{it}^{\text{total}}$ and, $pub.enroll_{it}^{\text{secondary}}$ so we can simply calculate $pub.enroll_{it}^{\text{primary}}$.

Public enrollment, 1918–1968: Data for public secondary enrollment and for total public enrollment are available biennially in the *Statistical Abstract of the United States* (even numbered years) from 1918–1968. In addition, data is also available in 1925, 1945, 1947, and 1949, 1955, and 1959. We log linearly interpolate any missing values from 1918–1968.

Public enrollment, 1969–2000: Data from 1969 to 2000 are annual, and come from NCES, *State Comparisons of Education Statistics: 1969–70 to 1996–97 (Snyder et al. 1998)*, as well as updates available from the NCES website.

Private enrollment data

Private enrollment, 1840–1916: Data for total private enrollments are available from various censuses, by state in 1840, 1850, 1860, 1870, 1890, 1910, and 1920. We log linearly interpolate between the decennial values listed above for any non-decennial years.

Data for total private secondary enrollments are available on an annual basis from 1899 to 1916 from the *Statistical Abstracts of the United States*. For these years, we are able to take the measure of total private enrollment above and subtract secondary enrollment to arrive at private elementary enrollment.

Prior to 1899, we observe total private enrollment, but do not observe the breakdown into elementary and secondary. However, we do observe national aggregates in 1890. Proceeding as we did above in the public case, we calculate:

$$pri.enroll_{it}^{\text{primary}} = pri.enroll_{it}^{\text{total}} \frac{\sum_j pri.enroll_{j,1890}^{\text{primary}}}{\sum_j pri.enroll_{j,1890}^{\text{total}}}, \quad t \leq 1890 \tag{33}$$

$$pri.enroll_{it}^{\text{secondary}} = pri.enroll_{it}^{\text{total}} - pri.enroll_{it}^{\text{primary}} \tag{34}$$

We also log linearly interpolate the secondary enrollment figures for 1891–1898 using the 1890 value (calculated directly above), and the 1899 figures.

Private enrollment, 1918–1968: Data for private secondary enrollment and total private enrollment are available biennially in *Statistical Abstracts of the United States* (even numbered years) from 1918 to 1940 and 1948 to 1968. Data is also available in 1925, 1947, and 1949, 1955, and 1959. We log linearly interpolate any missing values from 1918 to 1968.

Private enrollment, 1969–2000: For the years 1968–1980, 1991, 1993, 1995, 1997, and 1999, we observe private elementary and secondary enrollment figures from the *Digest of Education Statistic*s. We log linearly interpolate the 1992, 1994, 1996, and 1998 values.

For the years between 1980 and 1991, we are unable to obtain private elementary and private secondary enrollment figures by state directly. However we are able to obtain annual estimates of the national private elementary and private secondary totals from Projections of Education Statistics, various issues, as well as state-level data on Catholic elementary and Catholic secondary enrollment figures in 1985, 1988, and 1990–1999 from the *National Catholic Education Association*, various issues. We assume that the distribution of total private elementary and total private secondary enrollment figures across states is identical to the distribution of Catholic elementary and Catholic secondary enrollment figures across states. We inflate the Catholic state-level data enrollment data to correspond to the national totals for 1985, 1988, and 1990. We log linearly interpolate values for years 1981–1984, 1986–1987, and 1988.

Higher education enrollment data

1840–1899: Data for states are available from decennial census data in 1840, 1850, 1860, and 1870. In 1886, 1890, and 1891 data are available, typically subdivided into Medical, Theological, Law, and Liberal Arts enrollments. Data for non-census years between 1840 and 1870, as well as 1871–1885, 1887–1889, and 1892–1898 are log linearly interpolated.

1899–1920: Data are reported annually in *Statistical Abstracts* under a variety of titles and formats. Total higher education enrollment is the sum of sources below, except where enrollment figures are included in more than one source.

1. Schools of Technology and Institutions conferring only the B.S. degree (1899–1905)
2. Colleges and Seminaries for Women which confer degrees (1899–1910)
3. Coeducational Colleges and Universities and Colleges for men only (1899–1916, 1918)
4. Undergraduate Students in Univ., Colleges, and Schools of Tech. (1911–1916, 1918, 1920)
5. Professional Schools (1899–1916)
6. Public and Private Normal Schools (1899–1916, 1918, 1920)
7. Training Schools for Nurses, Comm. Schools, Manual and Industrial Training Schools (1910–1916, 1918, 1920)

1922–1946: Data is reported biennially in the *Statistical Abstracts* from 1922 to 1940, various issues, as Enrollment in Universities, Colleges, and Preparatory Schools. Similar data is also reported as Higher Education Enrollment in 1942, 1944, and 1946. Non-biennial years are log linearly interpolated.

1947–1968: Data is reported annually in *Statistical Abstracts*, various issues, as Institutions of Higher Educational, Fall Enrollment.

1969–2000: Data is reported in *State Comparisons of Education Statistics*. Higher educational enrollment is the sum of 2-year private, 2-year public, 4-year private, and 4-year public higher educational enrollment.

Population

We generally observe the age distribution of population in decennial years, beginning in 1840. In most cases, we are given data with 5-year population distributions. The usual structure is

<5, 5–9, 10–14, 15–19, 20–24…55–59, 60–64, 65–69, 70–74….

With the exception of calculating the average age of the population in a state, we are ultimately interested in the age groups: 5–13, 14–17, 18–24, 16–65. In order to calculate the number of persons in each group, we assume a uniform distribution of population within each age group.

In 1840, the white age distribution is reported, but only broad categories of the black age distribution are available. In order to allocate the total black distribution amongst the various age groups, we assume the fraction of total black population in each age group is identical to the fraction in the 1850 black distribution.

Labor force

All labor force data prior to 1970 is available at a decadal frequency. For non-decennial years prior to 1970, data is log linearly interpolated. Labor force data for 1840–1900 comes from Weiss (1999). Data for 1910–1940 is gainful workers, 10 years old and over, and is taken from *Historical Statistics of the United States: Colonial Times to 1970*, pp. 129–131. Data for 1950 and 1960 is decennial Census of Population data, and includes persons aged 14 and over. Data from 1970 to 2000 is Civilian Labor Force, 16 years and older, and is taken from the Bureau of Labor Statistics website.

Price levels

National price level data from 1875 to 1999 is the GDP deflator, as reported in Gordon, *Macroeconomics*, 7th edition, pp. A1–A3. National price level data prior to 1875 is the wholesale price index (all commodities) from Warren and Pearson, printed in *Historical Statistics of the United States: Colonial Times to 1970*, pp. 201–202. Data from 1840 to 1875 are normalized to correspond to the price level given by Gordon in 1875.

In addition, we use three sources of information on relative price levels across regions. Mitchener and McLean (1997) and Williamson and Linder (1980) provide regional price levels for census regions which we use from 1840 to 1960. Data from the two sources is primarily non-overlapping. Where we have data from both sources, we take the arithmetic average of the relative price level in each region. Prior to 1880 these sources do not include relative price levels for the Pacific and Mountain region. For data prior to 1880 in each of these two regions, we extrapolate the relative regional price level using the trend observed from 1880 to 1920. Berry et al. (2000) display price levels for each state on an annual basis from 1960 to 2000. To maintain consistency, we aggregate these state-level estimates into census regions. In non-decennial years, we interpolate relative price levels. We normalize regional price levels in all years to the national price level figures given in Gordon (and Warren and Pearson). All income measures are reported in 2000 dollars.

Expected years

The portion of the population, 25 years old and over that has completed various levels of school is given in the Census of the population in 1940–2000. From this information, we calculate the expected number of years of school completed, conditional on being in either the primary, secondary, or higher educational group. The values for yrs_t^{college}, $yrs_t^{\text{secondary}}$, and yrs_t^{primary} were obtained from decennial census data. Let $N(i-j)$ be the number of people who have completed between i and j years of schooling, inclusive.

$$yrs_{1940}^{\text{primary}} = \frac{2.5N(1-4) + 5.5N(5-6) + 7.5N(7-8)}{N(1-4) + N(5-6) + N(7-8)} \tag{35}$$

$$yrs_{1950,1960,1970,1980}^{\text{primary}} = \frac{2.5N(1-4) + 5.5N(5-6) + 7N(7) + 8N(8)}{N(1-4) + N(5-6) + N(7) + N(8)} \tag{36}$$

$$yrs_{1990}^{\text{primary}} = \frac{2.5N(1-4) + 7.23N(5-8)}{N(1-4) + N(5-8)} \tag{37}$$

$$yrs_{2000}^{\text{primary}} = \frac{6.42N(0-8)}{N(0-8)} \tag{38}$$

$$yrs^{\text{secondary}}_{1940,1950,1960,1970} = \frac{10N(9-11)+12N(12)}{N(9-11)+N(12)} \tag{39}$$

$$yrs^{\text{secondary}}_{1980} = \frac{9N(9)+10N(10)+11N(11)+12N(12)}{N(9)+N(10)+N(11)+N(12)} \tag{40}$$

$$yrs^{\text{secondary}}_{1990,2000} = \frac{10.5N(9-12)+12N(12)}{N(9-12)+N(12)} \tag{41}$$

$$yrs^{\text{college}}_{1940,1950,1960} = \frac{14N(13-15)+17N(16^{+})}{N(13-15)+N(16^{+})} \tag{42}$$

$$yrs^{\text{college}}_{1970} = \frac{14N(13-15)+16N(16)+18N(17^{+})}{N(13-15)+N(16)+N(17^{+})} \tag{43}$$

$$yrs^{\text{college}}_{1980} = \frac{\begin{array}{c}13N(13)+14N(14)+15N(15)+16N(16)\\ +17.5N(17-18)+20N(19^{+})\end{array}}{N(13)+N(14)+N(15)+N(16)+N(17-18)+N(19^{+})} \tag{44}$$

$$yrs^{\text{college}}_{1990} = \frac{\begin{array}{c}14N(scn+a)+16N(b)\\ +18N(ma)+19.75N(pr)+20N(d)\end{array}}{N(scn)+N(a)+N(b)+N(ma)+N(pr)+N(d)} \tag{45}$$

$$\tag{46}$$

$$yrs^{\text{college}}_{2000} = \frac{14N(sc)+14N(a)+16N(b)+18N(ma)+19.75N(prg)}{N(sc)+N(a)+N(b)+N(ma)+N(prg)} \tag{47}$$

9–12 = 9th to 12th grade, no diploma, sc = some college, scn = some college no degree, a = associate degree, b = bachelor's degree, ma = master's degree, prg = professional. or graduate degree, pr = professional school degree, d = doctorate degree.

In 1990, data are not reported as finely for those who have completed between 5 and 8 years of schooling. We need to assign a number of years of schooling to give to the group N(5-8), but this distribution is highly skewed. We calculate the conditional distribution in the years 1960, 1970, and 1980. We assign 7.23 years in 1990.

$$yrs^{5\text{–}8}_{1960} = \frac{5.5N(5-6)_{1960}+7N(7)_{1960}+8N(8)_{1960}}{N(5-6)_{1960}+N(7)_{1960}+N(8)_{1960}} = 7.22 \tag{48}$$

$$yrs^{5\text{–}8}_{1970} = \frac{5.5N(5-6)_{1970}+7N(7)_{1970}+8N(8)_{1970}}{N(5-6)_{1970}+N(7)_{1970}+N(8)_{1970}} = 7.23 \tag{49}$$

$$yrs^{5\text{–}8}_{1980} = \frac{5.5N(5-6)_{1980}+7N(7)_{1980}+8N(8)_{1980}}{N(5-6)_{1980}+N(7)_{1980}+N(8)_{1980}} = 7.24 \tag{50}$$

$$yrs^{5\text{–}8}_{1990} = 7.23 \tag{51}$$

In 2000, we need to assign a number of years of schooling to give to the group N(0–8), whose distribution is highly skewed. We use March 2000 CPS data for the population of people age 15 or over, which gives us data that is less aggregated than the census data. We assign 7.74

years to $N(7-8)$, which is the average value from the 1960 (7.73), 1970 (7.75), and 1980 (7.75) $yrs^{5\text{–}8}$. Thus the calculated value for $yrs^{0\text{–}8}_{2000}$ is 6.42:

$$yrs^{0\text{–}8}_{2000} = \frac{2.5N(1-4)_{2000} + 5.5N(5-6)_{2000} + 7.74N(7-8)_{2000}}{N(1-4)_{2000} + N(5-6)_{2000} + N(7-8)_{2000}} = 6.42 \qquad (52)$$

Values for yrs^{i}_t for periods prior to 1940 were calculated by log linearly interpolating from an initial value for the year in which the state first has adequate data available (see Table A1) to the 1940 value. Initial values are 4, 10, and 14 for primary, secondary, and higher education, respectively.

All values for non-census years between 1940 and 2000 were log linearly interpolated. We do not include those persons for whom the educational attainment level is not reported.

Description of calculations

Enrollment rates

Enrollment figures for public and private school are summed to obtain a total primary enrollment rate, total secondary enrollment rate, and total higher educational enrollment rate. From enrollment data, enrollment rates are calculated as below:

$$tot.enroll^{\text{primary}}_t = pub.enroll^{\text{primary}}_t + pri.enroll^{\text{primary}}_t \qquad (53)$$

$$tot.enroll^{\text{secondary}}_t = pub.enroll^{\text{secondary}}_t + pri.enroll^{\text{secondary}}_t \qquad (54)$$

$$tot.enroll^{\text{college}}_t = pub.enroll^{\text{college}}_t + pri.enroll^{\text{college}}_t \qquad (55)$$

$$r^{\text{primary}}_t = \frac{tot.enroll^{\text{primary}}_t}{\ell[5-13]_t} \qquad (56)$$

$$r^{\text{secondary}}_t = \frac{tot.enroll^{\text{secondary}}_t}{\ell[14-17]_t} \qquad (57)$$

$$r^{\text{college}}_t = \frac{tot.enroll^{\text{college}}_t}{\ell[18-24]_t} \qquad (58)$$

Educational exposure fractions

General methodology: To calculate the stock of human capital of each type, primary school stock, secondary school stock and higher education stock, we used a perpetual inventory method. The following will illustrate the nature of our calculations. We ignore state subscripts without loss of information. In period $t+1$, the stock of adults, with exposure to education level i, $i =$primary, secondary, and higher, but no more is given by

$$H^i_{t+1} = H^i_t(1-\delta^i_t) + I^i_t \qquad (59)$$

where δ^i_t is the departure rate from the labor force and I^i_t is the flow of new adults with exposure to education level i and no more. We first illustrate the general methodology where we assume a common departure rate for all education categories. We then estimate the departure rate separately for the secondary and higher educational classes.

It is useful to put the human capital measure as a fraction of the labor force. Thus, we normalize and produce

$$\frac{H_{t+1}^{i}}{L_{t+1}} = \frac{H_t^i}{L_t}\frac{L_t}{L_{t+1}}(1-\delta_t) + \frac{I_t^i}{L_{t+1}} \tag{60}$$

$$h_{t+1}^{i} = h_t^i\frac{L_t}{L_{t+1}}(1-\delta_t) + \frac{I_t^i}{L_{t+1}} \tag{61}$$

where h_t^i measures the share of the labor force exposed to education level i, and no more in year t. The flows into education categories are given by

$$I_t^{\text{college}} = \frac{r_t^{\text{college}}\Theta_t lfpr_t^{\text{college}}\ell[18-24]_t}{7} \tag{62}$$

$$I_t^{\text{secondary}} = \frac{\left(r_t^{\text{secondary}} - r_t^{\text{college}}\Theta_t\right) lfpr_t^{\text{secondary}}\ell[14-17]_t}{4} \tag{63}$$

$$I_t^{\text{primary}} = \frac{\left(r_t^{\text{primary}} - r_t^{\text{secondary}}\right) lfpr_t^{\text{primary}}\ell[5-13]_t}{9} \tag{64}$$

$$I_t^{\text{none}} = \frac{\left(1 - r_t^{\text{primary}}\right) lfpr_t^{\text{none}}\ell[5-13]_t}{9} \tag{65}$$

where r_t^{i} i = college, secondary and primary are the respective enrollment rates, $lfpr_t^{\text{i}}$ are the labor force participation rates for education category i, $\ell[i-j]$ is the number of people between the ages of i and j, inclusive, and Θ_t is the decade and state specific parameter to adjust the inflow into the higher educational category, described below.

In order to proceed we need a measure of δ_t, the departure rate of adults. As $L_{t+1} = L_t(1-\delta_t) + I_t^{\text{college}} + I_t^{\text{secondary}} + I_t^{\text{primary}} + I_t^{\text{none}}$, dividing through by L_{t+1} and then using definitions above, allows for the calculation of $\frac{L_t}{L_{t+1}}(1-\delta_t)$:

$$\frac{L_t}{L_{t+1}}(1-\delta_t) = 1 - \frac{\frac{r_t^{\text{college}}\Theta_t lfpr_t^{\text{college}}\ell[18-24]_t}{7} + \frac{\left(r_t^{\text{secondary}} - r_t^{\text{college}}\Theta_t\right)lfpr_t^{\text{secondary}}\ell[14-17]_t}{4} + \frac{\left(r_t^{\text{primary}} - r_t^{\text{secondary}}\right)lfpr_t^{\text{primary}}\ell[5-13]_t}{9} + \frac{\left(1-r_t^{\text{primary}}\right)lfpr_t^{\text{none}}\ell[5-13]_t}{9}}{L_{t+1}} \tag{66}$$

With this information, we can calculate each of the shares of the labor force in each schooling category.

Higher education/Higher education inflow adjustment (Θ): Using this method produced a much smaller share of the labor force exposed to higher education than the census figures. Thus we estimate the departure rate of those exposed to higher education independently. We assumed that there is no death, just retirement from the labor force after 45 years of work. The stock of adults exposed to higher education is then given as:

$$H_{t+1}^{\text{college}} = H_t^{\text{college}} - I_{t-45}^{\text{college}} + I_t^{\text{college}} \tag{67}$$

$$\frac{H_{t+1}^{\text{college}}}{L_{t+1}} = \frac{H_t^{\text{college}}}{L_t}\frac{L_t}{L_{t+1}} - \frac{I_{t-45}^{\text{college}}}{L_{t-45}}\frac{L_{t-45}}{L_{t+1}} + \frac{I_t^{\text{college}}}{L_{t+1}} \tag{68}$$

Thus, to calculate the higher education share in period t, we must measure $I_{t-45}^{\text{college}}/L_{t-45}$, which requires higher education enrollment data in period $t-45$. For the earlier portion of our sample, we do not observe enrollment rates early enough to make this calculation. Where necessary, we linearly interpolate between the 0 and the value of the higher education enrollment rate the first time it is observed. See Table B.3 for the years in which each state is first calculated and for the first time we observe higher educational enrollment figures. Unfortunately we do not observe L_{t-45} until we have 45 years of state data. We assume a constant labor force participation rate and use additional population data to calculate L_{t-45} .

There is an additional complication concerning the higher educational category. Since our calculations of the inflow to all categories are equal to the total enrollment across all ages in the category divided by the total population across all age in the category, they implicitly assume the enrollment rate is constant across ages within each education category. We are assuming that enrollment rate of 12-year olds is the same as the enrollment rate of 13-year olds, and more problematically, that the enrollment rate of 18-year olds is identical to the enrollment rate of 19-year olds.

To the extent that this assumption is erroneous, such that enrollment rates decrease with age within a category, the true the inflow in to the category will be understated. For an illustration, consider an extreme case. Suppose there are 700 students whose age distribution is uniform across ages 18–24. Suppose that 70 of the 100 persons aged 18 are enrolled in higher education, while no one above age 18 is enrolled (a 100% attrition rate between age 18 and age 19). As enrollment data is reported to us aggregated across ages, the data we would observe would be a higher educational enrollment rate of 10% (70 enrolled students and 700 college-aged students). This would seem to imply that only 10 percent of college-aged students were being exposed to some college. In fact, in this case 70% of all college aged students are being exposed to some college.

While this assumption is implicit in our calculations for the inflow to all of the educational categories, it is most troublesome where there is a high attrition rate between ages. While the attrition rate between 11th and 12th grade is greater than zero, it certainly is the case that the attrition rate between the first and second year of college is larger. As a result, we feel it is necessary to increase the inflow into the higher education category to address this issue, and as such we multiply the measured inflow by a factor we denote denote Θ_t,where this parameter is state specific and decade specific from 1940 to 2000.

We next describe the methodology to obtain the value of Θ_t for each state. Recall that the equations for the law of motion for the higher education category and the inflow into the higher education category are:

$$H_{t+1}^{\text{college}} = H_t^{\text{college}} - I_{t-45}^{\text{college}} + I_t^{\text{college}} \tag{69}$$

$$I_t^{\text{college}} = \frac{r_t^{\text{college}}\Theta_t lfpr_t^{\text{college}}\ell[18-24]_t}{7} \tag{70}$$

We observe the time path of the enrollment rate, labor force participation rate, college-aged population and the labor force. By iteratively substituting in the law of motion equation, one could solve for $h_{t+10}^{\text{college}}$as a function of the initial condition h_t^{college}, and the time path of the other observables. Therefore, if we knew the initial and terminal conditions $h_t^{\text{college}}$and $h_{t+10}^{\text{college}}$, we could solve for the value of Θ_t. The decennial censuses report the fraction of the labor force that has been exposed to higher education at the decadal frequency from 1940 to 2000, which we denote $\widetilde{h}_t^{\text{college}}$. To calculate the value of Θ_t for the 1940–1950 period, we use $\widetilde{h}_{1940}^{\text{college}}$for the initial condition and use $\widetilde{h}_{1950}^{\text{college}}$for the terminal condition, and then solve for Θ_t for each state. Similarly, to calculate the value of Θ_t for the 1950–1960 period, we utilize information on $\widetilde{h}_{1950}^{\text{college}}$ and $\widetilde{h}_{1960}^{\text{college}}$ and continue in the same fashion for the remaining decades. The interpretation of Θ_t would be the value of that Θ_t that is consistent with the census initial condition, the census terminal condition, and the enrollment rates and other observables.

For the period prior to 1940, we have no available decennial census data on higher education attainment. We choose a value of theta equal to 1.33 for all states, which is the labor force weighted average value across all states in census years. While somewhat arbitrary, the higher education enrollment rate is still quite small prior to 1940, only 8.2% for the nation as a whole in 1940. We experimented with alternative values for Θ including state specific Θ and this had little to no quantitative impact.

Secondary education/secondary departure rate (δ): After making the adjustments for higher educated category described above, we then utilized a common departure rate for the remaining educational categories (secondary, primary, and none). However, we found that this resulted in calculated shares exposed to elementary education that were less than zero in some states. As a result, we utilize a separate departure rate for the secondary category, $\delta_t^{\text{secondary}}$, and a departure rate for the remaining elementary and none categories, $\delta_t^{\text{primary}}$.

To determine the value of $\delta_t^{\text{secondary}}$ we proceed with the same general procedure as was utilized for the higher education category. We again observe the time path of enrollment rates, the labor force participation rates, the secondary-aged population and the labor force. By iteratively substituting into the law of motion equation, one could solve for $h_{t+10}^{\text{secondary}}$ as a function of the initial condition $h_t^{\text{secondary}}$, and the time path of the other observables. The result would be a 10th order polynomial in $(1 - \delta_t^{\text{secondary}})$, where we assumed that within the decade $\delta_t^{\text{secondary}}$ is constant. As with the higher education category, the census provides data at the decadal frequency from 1940 to 2000 on the fraction of the labor force that has been exposed to secondary education which we denote $\widetilde{h}_t^{\text{secondary}}$. To calculate $\delta_t^{\text{secondary}}$ from 1940 to 1950, we proxy for the initial condition using $\widetilde{h}_{1940}^{\text{secondary}}$. For each state and decade, we utilize a simple grid search. We begin with a value of $\delta_t^{\text{secondary}} = 0.0001$ and then increase the value in increments of 0.0001.[37] For each incremental value of $\delta_t^{\text{secondary}}$, we use the initial condition $\widetilde{h}_{1940}^{\text{secondary}}$and methodology described above to calculate the time path of the fraction of the labor force exposed to secondary education, $\widehat{h}_t^{\text{secondary}}$. We then compare, $\widehat{h}_{1950}^{\text{secondary}}$, the value in the terminal period implied by the initial condition and that specific value of delta, to the value reported by the decennial census, $\widetilde{h}_{1950}^{\text{secondary}}$. We choose the value of $\delta_t^{\text{secondary}}$ that most closely matches $\widehat{h}_{1950}^{\text{secondary}}$ to $\widetilde{h}_{1950}^{\text{secondary}}$. The interpretation of $\delta_t^{\text{secondary}}$ would be the value that is consistent with the initial census condition, the terminal census condition, and enrollment rates and other observables. We continue by utilizing the values of $\widetilde{h}_{1950}^{\text{secondary}}$ and $\widetilde{h}_{1960}^{\text{secondary}}$to calculate the value of $\delta_t^{\text{secondary}}$ from 1950 and 1960, and do the same for the remaining decades.

[37] In order to fit the shares of the labor force exposed to secondary school and no more, we allowed for the $1 - \delta^{\text{secondary}}$ term to exceed 1. While this would be problematic in an infinite horizon world, it is not for a 10 year horizon. The states where $1 - \delta^{\text{secondary}} > 1$ are those states with high rates of population growth, much of it driven by internal migration from other states of the US. For example the states with these unusual values, both labor force weighted 1940–2000 and unweighted 1940–2000, are: Florida, Arizona, Colorado, Nevada, and Alaska.

For values prior to 1940, use the value of delta calculated between 1940 and 1950, capped from above by .9999.

Elementary and no education: Having selected the value of $\delta_t^{\text{secondary}}$, we can calculate the share of workers exposed to secondary education using the following equation:

$$h_{t+1}^{\text{secondary}} = h_t^{\text{secondary}} \frac{L_t}{L_{t+1}} (1 - \delta_t^{\text{secondary}}) + \frac{I_t^{\text{secondary}}}{L_{t+1}} \tag{71}$$

Given that we have calculated for h_t^{college} and $h_t^{\text{secondary}}$ in all periods, we can proceed to calculate the shares for primary and no schooling.

The next set of equations shows how we can identify the term $\frac{L_t}{L_{t+1}} \left(1 - \delta_t^{\text{primary}}\right)$.

$$L_{t+1} = H_{t+1}^{\text{college}} + H_{t+1}^{\text{secondary}} + H_{t+1}^{\text{primary}} + H_{t+1}^{\text{none}} \tag{72}$$

$$L_{t+1} = H_{t+1}^{\text{college}} + H_{t+1}^{\text{secondary}} + \left(H_t^{\text{primary}} + H_t^{\text{none}}\right) \left(1 - \delta_t^{\text{primary}}\right) + \left(I_t^{\text{primary}} + I_t^{\text{none}}\right) \tag{73}$$

$$1 - h_{t+1}^{\text{college}} - h_{t+1}^{\text{secondary}} = \left(h_t^{\text{primary}} + h_t^{\text{none}}\right) \frac{L_t}{L_{t+1}} \left(1 - \delta_t^{\text{primary}}\right) + \frac{I_t^{\text{primary}} + I_t^{\text{none}}}{L_{t+1}}$$

$$\frac{L_t}{L_{t+1}} \left(1 - \delta_t^{\text{primary}}\right) = \frac{1 - h_{t+1}^{\text{college}} - h_{t+1}^{\text{secondary}} - \left(\frac{I_t^{\text{primary}}}{L_{t+1}} + \frac{I_t^{\text{none}}}{L_{t+1}}\right)}{\left(h_t^{\text{primary}} + h_t^{\text{none}}\right)} \tag{74}$$

$$\frac{L_t}{L_{t+1}} \left(1 - \delta_t^{\text{primary}}\right) = \frac{1 - h_{t+1}^{\text{college}} - h_{t+1}^{\text{secondary}} - \left(\begin{array}{c} \frac{\left(r_t^{\text{primary}} - r_t^{\text{secondary}}\right) lfpr_t^{\text{primary}} \ell[5-13]_t}{9} \\ + \frac{\left(1 - r_t^{\text{primary}}\right) lfpr_t^{\text{none}} \ell[5-13]_t}{9} \end{array}\right)}{\left(h_t^{\text{primary}} + h_t^{\text{none}}\right)} \tag{75}$$

We occasionally measure primary and secondary enrollment rates that are larger than unity. There are a couple of reasons why this occurs. The data contains individuals that were held back in school, and also there are people that receive education for the first time starting at an unusual age. Since we have very limited information on repeaters as well as unusual starters, we treat all cases as the latter.

Values of Θ_t and $1 - \delta_t^{\text{secondary}}$

Table B1a Average values of Θ_t labor force weighted and unweighted

NE	Θ^u	Θ^w	ESC	Θ^u	Θ^w	WNC	Θ^u	Θ^w
CT	1.694	1.473	AL	1.180	1.188	IA	1.099	1.099
ME	1.677	1.635	KY	1.119	1.146	KS	1.255	1.230
MA	1.078	1.033	MS	1.080	1.095	MN	1.511	1.544
NH	1.737	1.727	TN	1.337	1.359	MO	1.155	1.180
RI	1.042	0.958				NE	1.172	1.169
VT	1.217	1.212				ND	0.986	0.983
						SD	1.215	1.215
MA			WSC			ENC		
NJ	2.028	1.815	AR	1.254	1.301	IL	1.215	1.187
NY	1.000	0.977	LA	1.106	1.074	IN	1.176	1.175
PA	1.021	1.015	OK	0.961	0.985	MI	1.197	1.166
			TX	1.583	1.560	OH	1.216	1.191
						WI	1.270	1.238
SA			MTN			PAC		
DE	2.049	1.666	AZ	1.805	1.501	AK	2.112	1.879
DC	0.753	0.724	CO	1.812	1.768	CA	1.585	1.312
FL	2.574	2.054	ID	1.603	1.666	HI	1.906	1.653
GA	2.065	2.157	MT	1.435	1.470	OR	1.531	1.522
MD	1.691	1.552	NV	4.236	3.052	WA	1.717	1.680
NC	1.448	1.458	NM	1.731	1.461			
SC	1.567	1.560	UT	1.327	1.399			
VA	1.678	1.564	WY	1.514	1.351			
WV	0.672	0.670						

Table B1b Average values of $1 - \delta_t^{\text{secondary}}$ labor force weighted and unweighted

NE	$1 - \delta^{\mathrm{u}}$	$1 - \delta^{\mathrm{w}}$	ESC	$1 - \delta^{\mathrm{u}}$	$1 - \delta^{\mathrm{w}}$	WNC	$1 - \delta^{\mathrm{u}}$	$1 - \delta^{\mathrm{w}}$
CT	0.988	0.982	AL	0.981	0.982	IA	0.973	0.973
ME	0.981	0.982	KY	0.977	0.979	KS	0.982	0.981
MA	0.978	0.976	MS	0.972	0.973	MN	0.986	0.987
NH	1.000	1.000	TN	0.989	0.991	MO	0.984	0.985
R I	0.980	0.978				NE	0.977	0.978
VT	0.985	0.986				ND	0.957	0.958
						SD	0.970	0.972

Table B1b continued

MA			WSC			ENC		
NJ	0.988	0.985	AR	0.975	0.977	IL	0.983	0.981
NY	0.974	0.973	LA	0.977	0.979	IN	0.985	0.983
PA	0.975	0.975	OK	0.973	0.975	MI	0.983	0.982
			TX	0.998	0.998	OH	0.983	0.981
						WI	0.985	0.985
SA			MTN			PAC		
DE	0.996	0.992	AZ	1.030	1.023	AK	1.014	1.002
DC	0.962	0.961	CO	1.012	1.011	CA	1.003	0.992
FL	1.026	1.017	ID	0.990	0.994	HI	0.988	0.983
GA	0.998	1.003	MT	0.980	0.982	OR	1.000	0.999
MD	1.004	0.997	NV	1.050	1.045	WA	1.000	0.999
NC	0.988	0.992	NM	0.992	0.986			
SC	0.987	0.990	UT	0.997	0.998			
VA	0.998	0.996	WY	0.985	0.981			
WV	0.955	0.955						

The values are displayed in the Fig. B1 and B2 below. For brevity we present only the unweighted values by census region for both the $1 - \delta_t^{\text{sec ondary}}$ and the Θ_t, the weighted values look similar.

Initial conditions: The initial condition for h_t^i, $i =$ college, secondary and primary were the respective enrollment rate of each class divided by two.

Educational exposure fractions for foreign born

In the calculation of our measure of years of schooling in state i, recall that we multiply the fraction of state i's residents that were born in state j by the years of schooling in state j (assuming no mobility):

$$E_{it} = \sum_j S_{ijt} \widehat{E}_{jt} \tag{76}$$

We derived our measure of $\widehat{E}_{jt}$ from observing the enrollment rates in state j and using the perpetual inventory methodology described above. Because a fraction of the residents of state i's residents are foreign born, we require a measure of $\widehat{E}_{for,t}$, the average years of schooling for the foreign born. If we could observe the share of the foreign born in each education category, we would simply calculate:

$$\widehat{E}_{for,t} = h_{for,t}^{\text{primary}} yrs_{for,t}^{\text{primary}} + h_{for,t}^{\text{secondary}} yrs_{for,t}^{\text{secondary}} + h_{for,t}^{\text{college}} yrs_{for,t}^{\text{college}} \tag{77}$$

However, this data is not available, and thus we cannot calculate the corresponding measures of $h_{for,t}^{\text{primary}}$, $h_{for,t}^{\text{secondary}}$ and $h_{for,t}^{\text{college}}$.

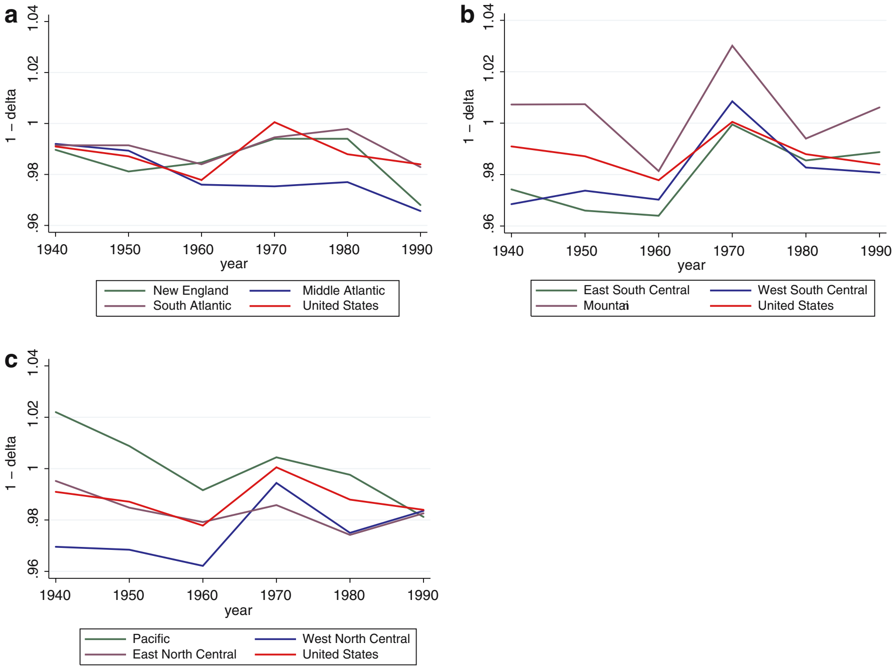

Fig. B1 Regional unweighted values of $1 - \delta_t^{\text{secondary}}$

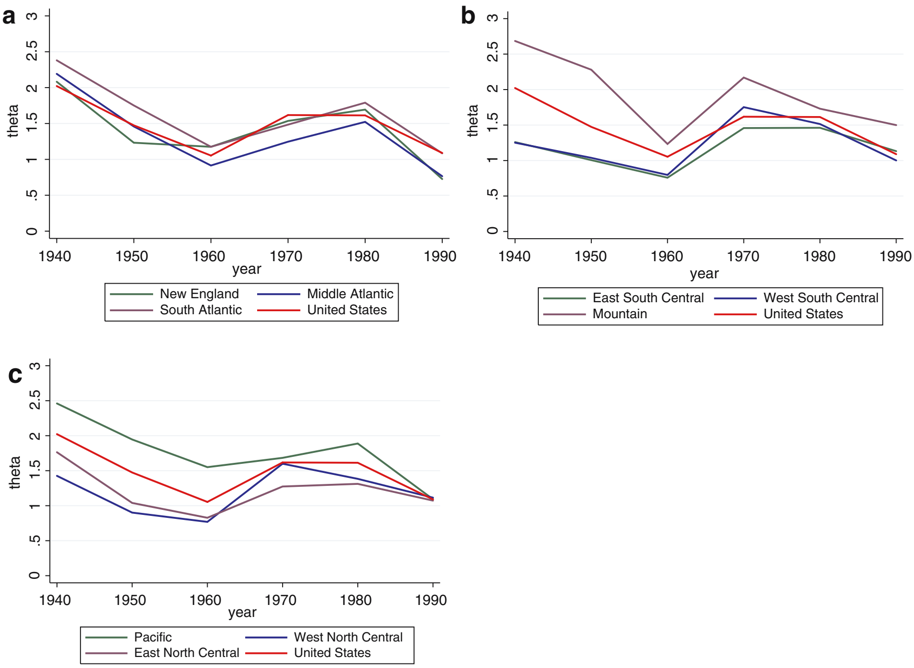

Fig. B2 Regional unweighted values of Θ_t

We use two different adjustment algorithms. We initially calculate the average years of schooling excluding the contributions made by the foreign born, which we denote $\widetilde{E}_{it}$:

$$\widetilde{E}_{it} = \sum_{j \neq for} S_{ijt} \widehat{E}_{jt} \tag{78}$$

We then assign the number of years of schooling to the foreign born $\widehat{E}_{for,t}$ so that our overall years of schooling measure, E_{it} equals the years of schooling reported by the census, yrscen_{it}:

$$\widehat{E}_{for,t} = \frac{\left(\mathrm{yrscen}_{it} - \widetilde{E}_{it}\right)}{S_{i,for,t}} \tag{79}$$

We then place a lower and upper bound on average years of schooling assigned to foreigners by

$$\widehat{E}_{for,t} \in \left[1, yrs_{it}^{\mathrm{college}}\right] \tag{80}$$

We allocate the shares among the educational categories such that:

$$\widehat{E}_{for,t} = \widehat{h}_{for,t}^{\mathrm{primary}} yrs_{it}^{\mathrm{primary}} + \widehat{h}_{for,t}^{\mathrm{secondary}} yrs_{it}^{\mathrm{secondary}} + \widehat{h}_{for,t}^{\mathrm{college}} yrs_{it}^{\mathrm{college}} \tag{81}$$

Although there is no unique allocation, we assigned the shares using the following algorithm, in order to preserve the equality of (81):

If $\widehat{E}_{for,t} < yrs_{it}^{\mathrm{primary}}$, we allocate between the none and primary categories, assigning zero for the secondary and college. In this case, $\widehat{E}_{for,t} = \frac{yrs_{it}^{\mathrm{primary}}}{S_{i,for,t}}$ and $\widehat{h}_{for,t}^{\mathrm{none}} = \left(1 - \widehat{h}_{for,t}^{\mathrm{primary}}\right)$. If $yrs_{it}^{\mathrm{primary}} < \widehat{E}_{for,t} < yrs_{it}^{\mathrm{secondary}}$, we assign zero for the none and college categories and allocate between the primary and secondary categories. If $yrs_{it}^{\mathrm{secondary}} < \widehat{E}_{for,t} < yrs_{it}^{\mathrm{college}}$, we assign zero for the none and primary categories and allocate between the secondary and college groups. If $\widehat{E}_{for,t} > yrs_{it}^{\mathrm{college}}$, we allocate between the secondary and college categories, assigning zero for the none and primary.

Idiosyncrasies

DC/MD/VA

We observe extremely high private enrollment rates for District of Columbia throughout the sample, presumably due to a large number of non-residents attending the District of Columbia schools. We surmise that these enrollment figures are overstated as many residents of Maryland and Virginia are attending District of Columbia schools. From 1910 to 1999, we assign a private elementary enrollment rate equal to zero for DC. We apportion those private elementary students enrolled in DC into the private elementary enrollment figures for Maryland and Virginia, using the population aged 5–13.

$$\begin{aligned} pri.enroll_{Md,t}^{\mathrm{primary}} &= pri.enroll_{Md,t}^{\mathrm{primary}} \\ &\quad + \left(\frac{\ell[5-13]_{Md,t}}{\ell[5-13]_{Va,t} + \ell[5-13]_{Md,t}}\right) pri.enroll_{DC,t}^{\mathrm{primary}} \end{aligned} \tag{82}$$

$$\begin{aligned} pri.enroll_{Va,t}^{\mathrm{primary}} &= pri.enroll_{Va,t}^{\mathrm{primary}} \\ &\quad + \left(\frac{\ell[5-13]_{Va,t}}{\ell[5-13]_{Va,t} + \ell[5-13]_{Md,t}}\right) pri.enroll_{DC,t}^{\mathrm{primary}} \end{aligned} \tag{83}$$

We allow the private secondary enrollment rate in DC to be no higher than the private secondary enrollment rate in the state of Massachusetts. We first calculate the enrollment rate in excess of the enrollment rate in DC, and then calculate the implied excess enrollment (students). We then apportion the excess enrollment into MD and VA, weighted by the population aged 14–17 in each state.

$$pri.enroll_{DC,t}^{\text{secondary}} = pri.r_{Ma,t}^{\text{secondary}} \ell[14-17]_{DC,t} \tag{84}$$

$$\begin{aligned} pri.enroll_{Md,t}^{\text{secondary}} &= pri.enroll_{Md,t}^{\text{secondary}} \\ &\quad + \left(\frac{\ell[14-17]_{Md,t} \cdot \left(pri.r_{DC,t}^{\text{secondary}} - pri.r_{Ma,t}^{\text{secondary}} \right)}{\ell[14-17]_{Va,t} + \ell[14-17]_{Md,t}} \right) \\ &\quad \times \ell[14-17]_{DC,t} \end{aligned} \tag{85}$$

$$\begin{aligned} pri.enroll_{Va,t}^{\text{secondary}} &= pri.enroll_{Va,t}^{\text{secondary}} \\ &\quad + \left(\frac{\ell[14-17]_{Va,t} \cdot \left(pri.r_{DC,t}^{\text{secondary}} - pri.r_{Ma,t}^{\text{secondary}} \right)}{\ell[14-17]_{Va,t} + \ell[14-17]_{Md,t}} \right) \\ &\quad \times \ell[14-17]_{DC,t} \end{aligned} \tag{86}$$

AK/HA

yrs_t^{college}, $yrs_t^{\text{secondary}}$, and yrs_t^{primary} for Alaska in 1939 and for Hawaii in 1940 were set as 14.5, 10.5, and 5.5, respectively.

ND/SD/Dakota

From 1880 through 1890, population and enrollment figures are reported for Dakota, which is the aggregate of North Dakota and South Dakota. In 1890, we first observe separate figures for North Dakota and South Dakota. Where data is available, we allocate a constant fraction of Dakota population and enrollment figures to each of North and South Dakota, based on the population of each state in 1890.

OK/Indian Territory

We first include Oklahoma in our data set only after the *Statistical Abstract* reported data for Oklahoma, rather than Indian Territory.

Labor's share of income

Table B2 Labor share and capital share of income[38]

line	period	*Emp.*	*Entrep.*	(1) + (2)	*Div.*(3)	*Int.*(4)	*Rent*(5)	(3) + (4) + (5)
		Comp.(1)	*Net Inc.*(2)					
1	1870–1880	50.0	26.4	76.5	15.8		7.8	23.6
2	1880–1890	52.5	23.0	75.4	16.5		8.2	24.6
3	1890–1900	50.4	27.3	77.7	14.7		7.7	22.4
4	1900–1910	47.1	28.8	75.8	15.9		8.3	24.2
5	1899–1908	59.5	23.8	83.3	5.3	5.1	6.4	16.7
6	1904–1913	59.6	23.3	82.9	5.7	5.1	6.3	17.1
7	1909–1918	59.7	23.3	83.0	6.5	4.9	5.7	17.0
8	1914–1923	63.0	20.8	83.8	5.6	5.3	5.3	16.2
9	1919–1928	65.1	18.3	83.4	5.4	6.0	5.2	16.6
10	1919–1928	61.7	19.5	81.2	5.6	6.1	7.1	18.8
11	1924–1933	63.1	16.6	79.7	6.5	7.8	5.9	20.3
12	1929–1938	64.9	15.9	80.8	6.6	8.4	4.3	19.2
13	1909–1913			69.5				30.5
14	1914–1918			67.0				33.0
15	1919–1923			69.5				30.5
16	1924–1928			69.7				30.3
17	1929–1933			69.2				30.8
18	1934–1938			70.4				29.6
19	1939–1943			72.1				27.9
20	1944–1948			74.9				25.1
21	1949–1953			74.5				25.5
22	1954–1958			77.3				22.7
23	1909–1958			71.4				28.6
24	1909–1929			68.9				31.1
25	1929–1958			73.0				27.0

Lines (13)–(25) from Table 4, Denison (1962) p. 30

[38] Lines (1)–(12) Table reprinted from Table 15, *National Income: A Summary of Findings*, Kuznets, NBER (1946), p. 50.

First year of data availability

Table B3 First year general enrollment data and higher education enrollment data is available

State	1st year general	1st year of higher ed.	State	1st year general	1st year of higher ed.
Alabama	1840	1840	Montana	1870	1870
Alaska	1939	1924	Nebraska	1860	1870
Arizona	1872	1899	Nevada	1870	1886
Arkansas	1840	1850	New Hampshire	1840	1840
California	1850	1860	New Jersey	1840	1840
Colorado	1870	1870	New York	1840	1840
Delaware	1840	1840	North Carolina	1840	1840
D.C.	1850	1850	North Dakota	1890	1890
Florida	1840	1870	Ohio	1840	1840
Georgia	1840	1840	Oklahoma	1890	1899
Hawaii	1940	1922	Oregon	1850	1860
Idaho	1870	1899	Pennsylvania	1840	1840
Illinois	1840	1840	Rhode Island	1840	1840
Indiana	1840	1840	South Carolina	1840	1840
Iowa	1840	1850	South Dakota	1890	1890
Kansas	1860	1860	Tennessee	1840	1840
Kentucky	1840	1840	Texas	1850	1850
Louisiana	1840	1840	Utah	1860	1870
Maine	1840	1840	Vermont	1840	1840
Maryland	1840	1840	Virginia	1840	1840
Massachusetts	1840	1840	Washington	1860	1870
Michigan	1840	1840	West Virginia	1870	1870
Minnesota	1860	1860	Wisconsin	1850	1850
Mississippi	1840	1840	Wyoming	1870	1890
Missouri	1840	1840			

Appendix C

To analyze the return to schooling, we need information on the income per worker. Since 1929, the Bureau of Economic Analysis has reported state-level annual income data. Total and per capita state income for 1840, 1880, 1900 and 1919–1921 are documented by Richard Easterlin in his works, "Interregional Differences in Per Capita Income, Population, and Total Income 1840–1950" in *Trends in the American Economy in the Nineteenth Century* and *Analyses of Economic Change in Population Redistribution and Economic Growth, United States, 1870–1950*. These data exclude transfer payments, likely small during this time period, and the figures for 1840 do not include all components of personal income. For the Census years not reported by Easterlin, 1850, 1860, 1870, 1890, and 1910, we generate the missing state per capita income using data available from the Easterlin sources above, the 1850 through

1910 Censuses, and the *Historical Statistics of the United States: Colonial Times to 1970* (HSUS). In order to calculate state per worker income, we calculate value added by each industry at the state level. Although data is not available for every industry, production value is reported for agriculture in the Census from 1870 to 1910 and production value and materials are reported in the Census from 1850 to 1910 for manufacturing.

Agricultural production value

From 1870 to 1910, each Census reports the value of agricultural products at the state level, Y_{it}^{ag}. To determine the state values of agricultural production for 1850, and 1860, we estimate the relationship of the production value of agricultural products sold within a state on the total value of farmland and buildings and agricultural labor force. We use national data at the decadal frequency from Towne and Rasmussen on the fraction of agricultural output that is value added.

Agricultural labor force is reported in the Census in 1840, 1850, and 1870 through 2000. There are two issues. The first, as documented by Weiss, suggests that the 1840 to 1870 and 1890 censuses systematically undersampled rural areas. As the labor force is likely to be almost exclusively engaged in agriculture in these areas, the census measure of agricultural labor is underestimated. Weiss provides an estimate of the overall labor force in each state, which we then compare to the overall labor force reported in the Census data. In cases where the Weiss estimate is larger than the census estimate, we interpret the difference as underreported agricultural labor and add this difference to the census measure of agricultural labor. Second, while the census does report a measure of the agricultural labor force in 1850, it usefulness is diminished because it does not include slave labor.[39] To estimate the total agricultural labor force for 1850 and 1860, we use the agricultural labor force reported in 1840, which includes slaves, and in 1870, which includes freed slaves, to construct the portion of the state labor force engaged in agricultural production, $\mathrm{fraction}_{it}^{ag}$. In non-slave holding regions, where the omission of slave labor is not problematic, we calculate $\mathrm{fraction}_{it}^{ag}$ in 1850 using the Census data.[40] We then linearly interpolate $\mathrm{fraction}_{it}^{ag}$ between 1840 and 1870 (between 1850 and 1870 for slave-holding regions and New England). We complete our measure of agricultural labor force in these intervening years by multiplying $\mathrm{fraction}_{it}^{ag}$ by the total labor force in each state.[41]

Values of agricultural products are not available in 1850, 1860, and 1920. We estimate the following relationship:

$$\ln\left(Y_{it}^{ag}\right) = \beta_1 \ln\left(farmvalue_{it}\right) + \beta_2 \ln\left(aglabor_{it}\right) + \beta_3 Z \tag{87}$$

To predict agricultural products 1850 to 1860, we use values from 1870 to 1880. To predict agricultural products in 1920, we estimate the relationship using data from 1910 and 1930.[42] The Census reports the production value of agricultural products and data on total farmland

[39] The 1860 census reports data hundreds of detailed occupations, but we do not attempt to map these occupations into the broader agricultural labor force.

[40] These regions are the Middle Atlantic, Mountain, Pacific, East North Central, and West North Central regions. We do not include the New England region because data in 1850 appear unreliable.

[41] No data on agricultural labor force is reported for Kansas, Nebraska, Texas, and Washington in 1840, therefore, we are unable to calculate the fraction of the labor force in agriculture using the methodology described above. For 1860, we proxy the agricultural labor force for these states by the number of persons listing their occupation as farmers.

[42] Additionally, data on agricultural products is not available in Arizona and New Mexico in 1890. We again regress using Eq. 88 and use data from 1880 and 1900 to estimate values for these two states.

Table C1 Regressions of natural log agricultural production

Variable	Coefficient	Std. error	Coefficient	Std. error
ln(farmvalue)	0.288	0.059	0.874	0.078
ln(aglabor)	0.577	0.066	0.147	0.080
NE	5.201	0.739	−0.812	0.968
MA	5.557	0.824	−0.936	1.055
SA	5.040	0.731	−0.898	1.000
ESC	5.281	0.759	−0.758	1.021
WSC	5.357	0.734	−0.878	1.043
MTN	5.028	0.613	−1.008	0.992
WNC	5.365	0.768	−1.189	1.102
ENC	5.482	0.817	−1.225	1.095
PAC	5.415	0.719	−1.174	1.078
N	86		96	
$\overline{R}^2$	0.9997		0.9997	
Data used	1870, 1880		1910, 1930	
Predict	1850,1860		1920	

value comes from HSUS. With our measures of agricultural capital, farmvalue$_{it}$, and labor, aglabor$_{it}$, where Z is the vector of region dummies and year$_t$ is a time trend. We then take the exponential of the predicted value, $\widehat{Y_{it}^{ag}}$, to estimate state level agricultural production value for 1850, 1860, and 1920. Results of these regressions are reported in Table C1.

We use national data at the decadal frequency from Towne and Rasmussen on the fraction of agricultural output that is value added to convert the predicted values into predicted value added agricultural output.

Manufacturing value added

The value added by manufacturers at the state level, Y$_{it}^{manu}$, is calculated by subtracting the value of materials used from the value of products sold reported in the Census from 1850 to 1920. Because the 1840 Census does not report the value added by manufacturing, we use the relationship between value added and the manufacturing labor force from 1850 to 1860 to determine value added in 1840. We regress the natural log of value added in the manufacturing sector, mvalue$_{it}$, on the natural log of the manufacturing labor force, mlabor$_{it}$, interacted with regions as well as individual census region effects, Z:[43]

$$\ln\left(mvalue_{it}\right) = \beta_1 Z + \beta_2\left(Z\ln\left(mlabor_{it}\right)\right) + \beta_3 year_t \tag{88}$$

Taking the exponential of the predicted $ln\ (\widehat{mvalue_{it}})$ generates the 1840 estimate of value added by manufacturing.

[43] Data on manufacturing labor are not available in 1890 and 1910. We calculate the fraction of the labor force engaged in manufacturing, fraction$_{it}^{min}$ in 1880, 1900, and 1920. We linearly interpolate the value of fraction$_{it}^{min}$ in 1890 and 1910, and multiply the result by the total labor force.

Mining value added

The output of precious metals is an important component of state income in the Pacific and Mountain region, particularly so in the early portion of out data set. As will be discussed in the following section, our income calculations allow for a component of income not captured by agriculture and mining. However, our methodology implicitly assumes that this component is relatively stable over time. Given the nature of gold and silver discoveries and subsequent rushes, we find this assumption unsatisfactory for these regions. As a result, we have collected data on precious metals mining output for the Mountain and Pacific regions.

Value added in the precious metals mining sector of the economy is calculated by subtracting the value of materials from the value of mining products, product_value$_{it}$, where available. A measure of mining products is available at the state level from the 1890 Census Report on Mineral Industries in the United States for 1870 and 1890.[44] A measure of materials used and labor is also available. This allows a measure of mining value added in 1890, $\mathrm{Y}^{mn}_{i,1890}$, to be calculated.

$$Y^{mn}_{i,1890} = product\_value_{it} - materials_{it} \tag{89}$$

We next calculate per worker value added in 1890:

$$y^{mn}_{i,1890} = \frac{Y^{mn}_{i,1890}}{L^{mn}_{i,1890}} \tag{90}$$

and fraction of output that is value added, $fracY_{i,1890}$:

$$fracY_{i,1890} = \frac{Y_{i,1890}}{product\_value_{i,1890}} \tag{91}$$

The 1870 Census report, The Statistics of Mining, gives data on employment, materials, and output of precious metals in 1870, but appears to be only a partial sample of all mining establishments. We do not use the measures of total products, value added and employment, but maintain measures of *per worker* products, value added, and employment.[45] Thus, we calculate $\mathrm{y}^{mn}_{i,1870}$ and $fracY_{i,1870}$ and then use these values with the 1890 values to interpolate to obtain $\mathrm{y}^{mn}_{i,1880}$ and $fracY_{i,1880}$. Prior to 1870, data is not as detailed. We assume that products per worker for each state in 1850 and 1860 is equal to it's value in 1870.[46] Thus:

$$y^{mn}_{i,1850} = y^{mn}_{i,1860} = y^{mn}_{i,1870} \tag{92}$$

We do the same for the fraction of products that is value added.

$$fracY^{mn}_{i,1850} = fracY^{mn}_{i,1860} = fracY^{mn}_{i,1870} \tag{93}$$

We next turn out attention to employment in precious metals mining. Direct measures of precious metals mining employment are available in 1840, and 1890 (and in 1870 we have a sample), as are measures of non-precious metal mining employment. This overlapping data will be exploited below. Data on precious metals employment data do not exist

[44] Data is not readily available from this source for 1890. Instead, we use the values in 1889.

[45] In addition, we maintain the fraction of all mining labor that is engaged in precious metals mining. See below.

[46] There is only one state, California, for which we have data in 1850. We make a separate adjustment for this state below.

directly in 1850, 1860, and 1880, yet measures of total employment in mining (precious and non-precious) are available in these years.

Let employment in precious metals mining be L_{it}^{prec}, and employment in non-precious metals mining, $L_{it}^{nonprec}$. In 1840, 1870, and 1890 we calculate:

$$fracL_{it}^{prec} = \frac{L_{it}^{prec}}{\left(L_{it}^{prec} + L_{it}^{nonprec}\right)} \tag{94}$$

For states in which we have no data prior to 1870, we assume that $fracL_{it}^{prec}$ in 1850 and 1860 are identical to the 1870 values in each state. We also interpolate between 1870 and 1890 to acquire 1880 values. Thus:

$$fracL_{i,1850}^{prec} = fracL_{i,1860}^{prec} = fracL_{i,1870}^{prec} \tag{95}$$

Next, we calculate labor in the precious metal sector, L_{it}^{prec}, in 1850, 1860, and 1880 as,

$$L_{it}^{prec} = fracL_{it}^{prec}\left(L_{it}^{prec\&nonprec}\right) \tag{96}$$

And to correct for the fact that L_{it}^{prec} in 1870 is a sample, we geometrically interpolate between the value of L_{it}^{prec} in 1860 and 1880.

Finally, we can calculate our measure of Y_{it}^{mn} for 1850, 1860, 1870, and 1880:

$$Y_{it}^{mn} = y_{it}^{mn} L_{it}^{mn} fracL_{it}^{prec} \tag{97}$$

As a check on the reasonableness of our calculations, we compare the sum of mining output across the states to the national output figures given for 1850 and 1860 in the 1890 Census report. We find we overestimate mining output in 1860. We assume that California has the same share of national mining output in 1860 as it does in 1850. We then renormalize all other states so that the sum is equal to the national total.

Total state income

Adding the value added of products produced by manufacturers and mines and the estimated value added from agricultural production at the state level generates the total state income attributable to manufacturing, mining, and agriculture:

$$Y_{it}^{ag+manu+mn} = Y_{it}^{ag} + Y_{it}^{manu} + Y_{it}^{mn} \tag{98}$$

for $1840 \leq t \leq 1920$.[47]

Unfortunately for us, this measure of income is not the total state income, but only the of portion of state income resulting from manufacturing, mining, and agriculture. In order to account for the remaining industries in a states' economy, we turn to the total income calculations reported by Easterlin. In *Trends in the American Economy in the Nineteenth Century*, Easterlin calculates the total state income level for 1840 and in *Analyses of Economic Change in Population Redistribution and Economic Growth, United States, 1870–1950,* he reports total state income for 1880, 1900, and 1919–1921(1920). For 1840, 1880, 1900, and 1920,

[47] We only make our mining adjustments in 1850, 1860, 1870, and 1890 for the Mountain and Pacific regions. We do not adjust mining for states outside of these regions. That is, $Y_{it}^{mn} = 0$ for all other regions.

we calculate the difference between our estimated, $Y_{it}^{ag+manu+mn}$, and Easterlin's total state income, Y_{it}^{E}:

$$Y_{it}^{not} = Y_{it}^{E} - Y_{it}^{ag+manu+mn} \tag{99}$$

for $t = 1840$, 1880, 1900, and 1920. We then calculate the ratio of income generated outside agriculture, manufacturing, and mining over income produced by agriculture, manufacturing, and mining:[48]

$$Y_{it}^{notshare} = \frac{Y_{it}^{not}}{Y_{it}^{ag+manu+mn}} \tag{100}$$

For the states with 1840 Easterlin incomes, listed in Table C2, we estimate the ratio of income generated outside agriculture, manufacturing, and mining over income produced by agriculture, manufacturing, and mining for 1850, 1860, 1870, 1890, and 1910 using the following methods:

$$\widehat{Y}_{i,1850}^{notshare} = \left(Y_{i,1840}^{notshare}\right)^{.75}\left(Y_{i,1880}^{notshare}\right)^{.25} \tag{101}$$

$$\widehat{Y}_{i,1860}^{notshare} = \left(Y_{i,1840}^{notshare}\right)^{.5}\left(Y_{i,1880}^{notshare}\right)^{.5} \tag{102}$$

$$\widehat{Y}_{i,1870}^{notshare} = \left(Y_{i,1840}^{notshare}\right)^{.25}\left(Y_{i,1880}^{notshare}\right)^{.75} \tag{103}$$

$$\widehat{Y}_{i,1890}^{notshare} = \left(Y_{i,1880}^{notshare}\right)^{.5}\left(Y_{i,1900}^{notshare}\right)^{.5} \tag{104}$$

$$\widehat{Y}_{i,1910}^{notshare} = \left(Y_{i,1900}^{notshare}\right)^{.5}\left(Y_{i,1920}^{notshare}\right)^{.5} \tag{105}$$

For the states without 1840 incomes, listed in Table C3, we use the 1880 ratio of income generated outside agriculture, manufacturing, and mining over income produced by agriculture, manufacturing, and mining, $Y_{i,1880}^{notshare}$, in order to determine $Y_{i,t}^{notshare}$, for $t = 1850, 1860, 1870$. For 1890, and 1910 we use the similar method as above:

$$\widehat{Y}_{i,1850}^{notshare} = \left(Y_{i,1880}^{notshare}\right) \tag{106}$$

$$\widehat{Y}_{i,1860}^{notshare} = \left(Y_{i,1880}^{notshare}\right) \tag{107}$$

$$\widehat{Y}_{i,1870}^{notshare} = \left(Y_{i,1880}^{notshare}\right) \tag{108}$$

$$\widehat{Y}_{i,1890}^{notshare} = \left(Y_{i,1880}^{notshare}\right)^{.5}\left(Y_{i,1900}^{notshare}\right)^{.5} \tag{109}$$

$$\widehat{Y}_{i,1910}^{notshare} = \left(Y_{i,1900}^{notshare}\right)^{.5}\left(Y_{i,1920}^{notshare}\right)^{.5} \tag{110}$$

Using these ratios we calculate our final total state income, $\widehat{Y_{it}^{all}}$, for all non-Easterlin years:

$$\widehat{Y}_{it}^{all} = Y_{it}^{ag+manu+mn}\left[1 + \widehat{Y}_{i,t}^{notshare}\right] \tag{111}$$

[48] We occasionally observe a measure of Y_{it}^{not} that is less than zero in 1840. For these states, the sum of agricultural, mining, and manufacturing income exceeds the figure given as total income by Easterlin. We replace the measure of Y_{it}^{not} with zero. Cases are rare and magnitudes are small.

Table C2 1840 State incomes reported by Easterlin

Alabama
Arkansas
Connecticut
Delaware
Florida
Georgia
Illinois
Indiana
Iowa
Kentucky
Louisiana
Maine
Maryland
Massachusetts
Michigan
Mississippi
Missouri
New Hampshire
New Jersey
New York
North Carolina
Ohio
Pennsylvania
Rhode Island
South Carolina
Tennessee
Vermont
Virginia
Wisconsin

Table C3 1840 State incomes not reported by Easterlin (with first year of agriculture and manufacturing data availability)

State	First year calculated	State	First year calculated
Arizona	1870	New Mexico	1850
California	1850	Oregon	1850
Colorado	1870	South Dakota	1910
Idaho	1870	Texas	1850
Kansas	1860	Utah	1850
Minnesota	1860	Washington	1860
Montana	1870	West Virginia	1870
Nebraska	1860	Wyoming	1870
Nevada	1870		

In order of find our calculated per worker income, we simple take total state income in year and divide it by the states' labor force reported by the census, except 1850 and 1860 where the our labor force figures are adjusted for slaves:

$$y_{it} = \frac{\widehat{Y}_{it}^{all}}{L_{it}} \tag{112}$$

We then put our per worker income measures into real terms by adjusting for both national and regional differences in prices. See Appendix B for more details on price levels.

Income bounds

In this section we present income per worker bounds for the period 1840–1920. Since we imputed non-agricultural, non-manufacturing, non-mining output per worker (the not sectors) for each state, we provide bounds on our estimates in this section. First, our procedure replaces our state level estimate of the non-agricultural, non-manufacturing, and non-mining output per worker with the national 10th (90th) percentile values of non-agricultural, non-manufacturing, non-mining output per worker. We then recalculate the overall real output per worker using these national 10th (90th) percentile bounds for the non-agricultural, non-manufacturing, and non-mining sector. The results do not change substantively if instead we use the census regions or the North, South and West region values instead. The figure below presents the results of this exercise. In the panel graph each panel presents our estimates of

Table C4 National and regional income and bounds (10th, 90th) percentiles

Region	1840	1850	1860	1870	1880	1890	1900	1910	1920	Average
US	4114	6691	7302	7612	9449	11514	11477	12554	14429	
	(.884, 1.113)	(.788, 1.443)	(.740, 1.408)	(.728, 1.248)	(.755, 1.103)	(.759, 1.171)	(.738, 1.147)	(.855, 1.170)	(.857, 1.188)	(.789, 1.221)
NE	5267	9077	9999	9717	10998	13818	13073	14230	15706	
	(.756, 1.067)	(.700, 1.123)	(.669, 1.114)	(.694, 1.082)	(.817, 1.125)	(.820, 1.215)	(.781, 1.163)	(.889, 1.177)	(.887, 1.175)	(.779, 1.138)
MATL	5528	7901	8840	10910	12954	16786	14947	16234	18469	
	(.885, 1.054)	(.768, 1.599)	(.688, 1.447)	(.643, 1.120)	(.693, 1.023)	(.691, 1.074)	(.676, 1.055)	(.798, 1.100)	(.796, 1.099)	(.733, 1.175)
SATL	2342	3302	3647	3728	4751	5400	5929	7909	9770	
	(.935, 1.322)	(.877, 1.579)	(.831, 1.637)	(.808, 1.570)	(.960, 1.519)	(.930, 1.557)	(.908, 1.560)	(.936, 1.293)	(.920, 1.310)	(.901, 1.483)
ESC	3683	5344	5928	4869	5447	5695	5900	6774	7947	
	(.965, 1.053)	(.931, 1.243)	(.879, 1.265)	(.850, 1.314)	(.880, 1.258)	(.854, 1.256)	(.898, 1.408)	(.952, 1.332)	(.973, 1.452)	(.909, 1.287)
WSC	5042	7392	7729	6312	5971	6922	7641	8633	11512	
	(.739, 1.001)	(.636, 1.218)	(.605, 1.302)	(.597, 1.043)	(.807, 1.211)	(.772, 1.174)	(.755, 1.227)	(.850, 1.149)	(.780, 1.115)	(.727, 1.161)
MTN	–	10250	12606	12124	10951	13840	13838	11789	13823	
		(.220, 1.000)	(.237, 1.000)	(.404, 1.079)	(.594, 1.085)	(.565, 1.033)	(.560, 1.018)	(.771, 1.144)	(.784, 1.160)	(.517, 1.065)
PAC	–	43207	24257	16500	13786	15438	14992	14188	17607	
		(.337, 1.000)	(.344, 1.000)	(.416, 1.000)	(.576, 1.000)	(.614, 1.057)	(.578, 1.012)	(.765, 1.109)	(.724, 1.045)	(.544, 1.028)
WNC	3503	4635	5698	6799	9248	10972	12395	13167	13486	
	(1.000, 1.259)	(1.000, 2.548)	(.862, 1.957)	(.837, 1.415)	(.743, 1.053)	(.784, 1.195)	(.756, 1.098)	(.907, 1.220)	(.937, 1.327)	(.870, 1.452)
ENC	4540	7335	7444	7288	11147	12965	13440	14682	15842	
	(.998, 1.148)	(.939, 1.647)	(.908, 1.599)	(.895, 1.452)	(.749, 1.045)	(.791, 1.183)	(.763, 1.126)	(.886, 1.167)	(.917, 1.215)	(.871, 1.287)

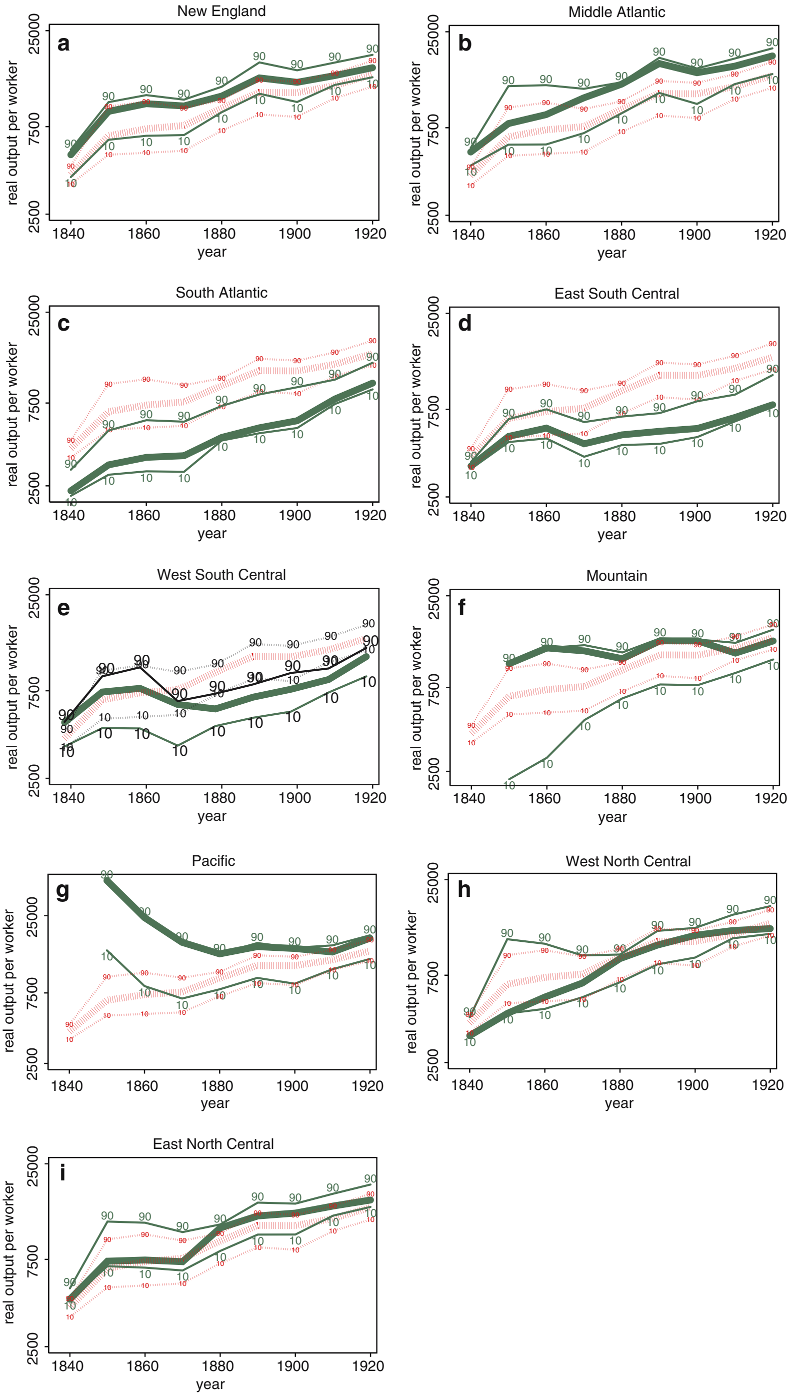

Fig. C1 Real output per worker: upper and lower bounds, log scale. *Note*: The Pacific region estimates of real output per worker for those not working in agriculture, manufacturing, or mining (the not sectors) are much higher than other regions during the early periods from 1850 to 1860

the region's output per worker, and the two bounds as well as the US values for each. The census region figures are always in green and the US values are always in red.

Table C4 presents our regional estimates as well as the percent deviation between the estimates and the two bounds for each census year.

These bounds are constructed using the following method. First we calculate the total income produced by workers not in agriculture, manufacturing, mining, $\widehat{Y}_{it}^{not}$, by state, year:

$$\widehat{Y}_{it}^{not} = \widehat{Y}_{it}^{all} - Y_{it}^{ag+man+mn} \tag{113}$$

We then calculate the per worker income for workers not in agriculture, manufacturing, mining, $\widehat{y}_{it}^{not}$, by dividing the total income produced by workers not employed in agriculture, manufacturing, mining by the number of workers in these other industries:

$$\widehat{y}_{it}^{not} = \frac{\widehat{Y}_{it}^{not}}{L_{it}^{not}} \tag{114}$$

In order to compare not per worker incomes across states, we deflate our nominal measures using the same deflator constructed in Appendix B (Price Levels), to create the real not per worker income, $\widetilde{y}_{it}^{not}$. For each state, we generate two real income per worker bound series: one by replacing a state's own real not per worker income with the national 10th percentile real not per worker income, multiplying by the number of workers in the not sector, adding this result to the total income from agriculture, mining, and manufacturing, and dividing by the total number of workers in the state:

$$\widetilde{y}_{it}^{10th} = \frac{\widetilde{y}_{it}^{10th,not} L_{it}^{not} + \widetilde{y}_{it}^{ag+man+mn} L_{it}^{ag+man+mn}}{L_{it}^{not} + L_{it}^{ag+man+mn}} \tag{115}$$

and the other by replacing a state's own real not per worker income with the national 90th percentile real per worker income for employees in the non-sector and repeating the similar process as above:

$$\widetilde{y}_{it}^{90th} = \frac{\widetilde{y}_{it}^{90th,not} L_{it}^{not} + \widetilde{y}_{it}^{ag+man+mn} L_{it}^{ag+man+mn}}{L_{it}^{not} + L_{it}^{ag+man+mn}} \tag{116}$$

For the 90th percentile, if a states actual not per worker income is higher; we simply used the states own. For the 10th percentile: if a state's real not per worker income was lower, we simply used the states own, so that for both series a state's overall real per worker income always lay on or between the constructed 90th and 10th percentile real income per worker bounds.

Appendix D

Table D1 below presents the labor force weighted correlations of our years of schooling in the labor force with the two separate state human capital measures of Mulligan and Sala-i-Martin (1997, 2000).

One way to compare our estimates of years of schooling in the labor force with the values of years of schooling by state from the Census is to compare the means and standard deviations, both weighted and unweighted. Table D2 provides evidence that our estimates are similar, if not identical with the census values. The largest differences occur in 1960. With the exception of 1960, the mean of our estimates differs from the Census by less than 1.1%. Our standard deviations closely match the census standard deviations.

Table D1 Correlation of years of schooling in the labor force with Mulligan and Sala-i-Martin (1997, 2000)

	Yrs of schooling	hc1997	hc2000
1940			
Yrs of schooling	1		
hc1997	.9258	1	
hc2000	.9138	0.9754	1
1950			
Yrs of schooling	1		
hc1997	.9306	1	
hc2000	.8851	.9311	1
1960			
Yrs of schooling	1		
hc1997	.8268	1	
hc2000	.8041	.9426	1
1970			
Yrs of schooling	1		
hc1997	.8326	1	
hc2000	.7567	.8639	1
1980			
Yrs of schooling	1		
hc1997	.8828	1	
hc2000	.7887	.9231	1
1990			
Yrs of schooling	1		
hc1997	.7835	1	
hc2000	.6521	.9418	1

Table D2 Average years of schooling: census and estimates

Year	Census mean	Census std. dev.	Estimate mean	Estimate std. dev.	% dev. mean	Census weighted mean	Estimate weighted mean	% dev. weighted mean
1940	8.48	1.07	8.51	1.03	0.3	8.37	8.41	0.5
1950	9.33	1.00	9.32	0.93	−0.1	9.32	9.33	0.0
1960	10.47	0.62	10.16	0.73	−3.0	10.46	10.23	−2.1
1970	10.97	0.65	10.85	0.63	−1.1	10.92	10.87	−0.5
1980	12.06	0.55	11.95	0.54	−0.9	12.01	11.96	−0.4
1990	12.82	0.42	12.80	0.44	−0.2	12.75	12.74	−0.1
2000	13.54	0.35	13.47	0.42	−0.5	13.52	13.48	−0.3

Table D3 Regressions of average years of schooling from the census on estimates (standard errors)

Variable	ALL	1940	1950	1960	1970	1980	1990	2000
Weighted								
E	0.9930	1.000	1.019	0.7280	1.003	0.965	0.965	0.877
	(0.005)	(0.032)	(0.032)	(0.050)	(0.034)	(0.035)	(0.033)	(0.041)
Constant	0.129	−0.046	−0.176	3.01	0.017	0.469	0.459	1.70
	(0.061)	(0.274)	(0.299)	(0.512)	(0.373)	(0.419)	(0.423)	(0.549)
N	355	49	51	51	51	51	51	51
ρ	.9897	.9617	.9513	.8853	.9415	.9415	.9547	.9359
prob $> F$	.0000	.3199	.8408	.0000	.0195	.0077	.3697	.0002
Unweighted								
E	0.9956	0.9957	1.031	0.7480	0.9727	0.9584	0.9151	0.785
	(0.008)	(0.041)	(0.047)	(0.055)	(0.049)	(0.048)	(0.040)	(0.042)
Constant	0.139	0.009	−0.280	2.87	0.419	0.608	1.11	2.97
	(0.085)	(0.351)	(0.443)	(0.565)	(0.534)	(0.579)	(0.516)	(0.562)
N	355	49	51	51	51	51	51	51
$\overline{R}^2$	.9795	.9249	.9049	.7837	.8866	.8865	.9115	.8760
prob $> F$	0.0000	.8000	.7868	.0000	.0009	.0003	.0488	.0000

An alternative way to compare our estimates with the census data is to regress the census years of schooling on our calculated years of schooling. Table D3 details how well we fit the census information using labor force weighted regressions as well as unweighted regressions.[49] Overall, the our calculations fit the data extremely well, but this may be a result of the observed time trend in education. Therefore, we also present results for each decade. If our estimates were identical to the census measures, the resulting slope coefficient of years of schooling would equal 1 and the intercept would equal 0. The final row of the table contains the result of the joint test of this hypothesis. Overall we reject the null hypothesis that our estimated slope coefficient is 1 and our intercept is 0, however for 1940, 1950, and 1990 (weighted) we cannot reject the null. Our fit is quite good, in the unweighted regressions our $\overline{R}^2$ are typically over .85 with the exception of 1960. For the weighted regressions, we report the correlation coefficient in the row marked ρ, as reported $\overline{R}^2$ are not meaningful in weighted regressions. With the exception of 1960, all correlations easily exceed .9.

To further determine the robustness of our methodology, we also compare the individual educational category components (workers exposed to elementary school and no more; workers exposed to secondary school and no more; and workers exposed to higher education) to those reported by the Census.[50] Tables D4 through D7 present the unweighted and labor force weighted means and standard deviations reported by the census as well as our calculated shares of the labor force represented by each education category.[51]

[49] This seems reasonable as it seems much more important to fit New York or California than to give those states equal weight as states like North and South Dakota.

[50] We thank the anonymous referees for suggesting this additional measure of goodness-of-fit.

[51] For those with no education exposure we match the means and standard deviations quite well, although the shares are so small that we choose not to report the regressions on these shares.

Table D4 Exposed to no schooling: census and estimates

Year	Census mean	Census std. dev.	Estimate mean	Estimate std. dev.	Census weighted mean	Estimate weighted mean
1940	.032	.023	.034	.035	.034	.032
1950	.022	.018	.027	.030	.022	.020
1960	.000	.000	.015	.017	.000	.012
1970	.012	.006	.013	.013	.012	.009
1980	.006	.004	.008	.009	.007	.006
1990	.007	.004	.006	.006	.009	.006
2000	.000	.000	.002	.003	.000	.001

Table D5 Exposed to elementary school and no more: census and estimates

Year	Census mean	Census std. dev.	Estimate mean	Estimate std. dev.	Census weighted mean	Estimate weighted mean
1940	.482	.076	.527	.115	.490	.554
1950	.378	.083	.415	.102	.380	.431
1960	.308	.074	.326	.082	.307	.324
1970	.208	.058	.259	.063	.206	.257
1980	.128	.042	.164	.048	.128	.164
1990	.065	.024	.090	.038	.065	.092
2000	.047	.016	.059	.033	.053	.063

Table D6 Exposed to secondary school and no more: census and estimates

Year	Census mean	Census std. dev.	Estimate mean	Estimate std. dev.	Census weighted mean	Estimate weighted mean
1940	.346	.062	.341	.093	.344	.320
1950	.423	.062	.414	.079	.429	.411
1960	.483	.048	.461	.063	.489	.462
1970	.525	.037	.496	.046	.532	.504
1980	.503	.039	.477	.036	.505	.479
1990	.437	.051	.430	.037	.433	.438
2000	.397	.053	.382	.040	.391	.378

Looking at the overall trend, our estimates are typically above the Census means for those exposed to elementary school and at or below the Census means for those exposed to secondary and higher education. This trend may be caused by students attending grade levels that do not correspond to their age. Recall that primary enrollment rates are calculated as the number of elementary students enrolled divided by the population of students that are typically elementary school aged: 5–13 years old. Age distribution data is not available for

Table D7 Exposed to higher education: census and estimates

Year	Census mean	Census std. dev.	Estimate mean	Estimate std. dev.	Census weighted mean	Estimate weighted mean
1940	.140	.033	.099	.038	.132	.094
1950	.177	.039	.144	.039	.169	.139
1960	.209	.039	.197	.037	.204	.202
1970	.255	.046	.232	.043	.249	.229
1980	.363	.060	.350	.053	.359	.351
1990	.491	.066	.474	.057	.493	.464
2000	.556	.060	.557	.054	.555	.559

Table D8 Regressions of exposed to elementary school and no more from the census on estimates (standard errors)

Variable	ALL	1940	1950	1960	1970	1980	1990	2000
Weighted								
E	0.8915	0.4515	0.7328	0.8430	0.9054	0.8286	0.4780	0.4814
	(0.011)	(0.083)	(0.081)	(0.071)	(0.073)	(0.072)	(0.058)	(0.078)
Constant	−0.009	0.240	0.064	0.034	-0.027	-0.008	0.021	0.023
	(0.003)	(0.047)	(0.035)	(0.024)	(0.019)	(0.012)	(0.006)	(0.005)
N	355	49	51	51	51	51	51	51
ρ	.9649	.7216	.8118	.8703	.8725	.8431	.7549	.6024
prob $>F$	.0000	.0000	.0000	.0003	.0000	.0000	.0000	.0000
Unweighted								
E	0.8963	0.4803	0.6614	0.7906	0.8043	0.7544	0.4680	0.2895
	(0.013)	(0.066)	(0.067)	(0.063)	(0.064)	(0.068)	(0.057)	(0.053)
Constant	−0.005	0.229	0.104	0.050	.0002	0.004	0.023	0.030
	(0.004)	(0.035)	(0.029)	(0.021)	(0.017)	(0.012)	(0.006)	(0.004)
N	355	49	51	51	51	51	51	51
$\overline{R}^2$	.9310	.5207	.6591	.7574	.7612	.7108	.5699	.3630
prob $>F$	.0000	.0000	.0000	.0001	.0000	.0000	.0000	.0000

any educational category. Therefore, there are two possibilities that may inflate the estimated portion of students exposed to elementary schooling: late starters and repeaters. For example, if a student begins formal schooling after age five and continues on to secondary schooling, he or she will be over 13 years of age while attending elementary school. In this case, this student is included in the numerator of the elementary enrollment rate, but not the denominator. Similarly, repeaters enrolled in elementary school when their age cohort is assumed to have finished may also serve to inflate the elementary exposure rates. Both cause elementary enrollment rates in excess of 100% and result in exposure estimates above those reported by the Census. Even with these data constraints, we find that our calculated means and standard deviations of our elementary shares to be close to the Census data.

Table D9 Regressions of exposed to secondary school and no more from the census on estimates (standard errors)

Variable	ALL	1940	1950	1960	1970	1980	1990	2000
Weighted								
E	0.9128	0.4871	0.6624	0.5754	0.5296	0.8415	1.017	1.127
	(0.026)	(0.085)	(0.078)	(0.073)	(0.069)	(0.087)	(0.123)	(0.080)
Constant	0.054	0.188	0.157	0.223	0.265	0.102	−0.012	−0.034
	(0.012)	(0.028)	(0.032)	(0.034)	(0.035)	(0.042)	(0.054)	(0.031)
N	355	49	51	51	51	51	51	51
ρ	.8507	7174	.7833	.7719	.7133	.6515	.7144	.8170
prob $> F$	.0000	.0000	.0000	.0000	.0000	.0000	.6182	.0005
Unweighted								
E	0.8425	0.4813	0.6209	0.5888	0.5891	0.7095	0.9947	1.105
	(0.028)	(0.067)	(0.069)	(0.068)	(0.081)	(0.115)	(0.137)	(0.110)
Constant	0.084	0.182	0.166	0.212	0.233	0.165	0.009	−0.025
	(0.012)	(0.024)	(0.029)	(0.032)	(0.040)	(0.055)	(0.059)	(0.042)
N	355	49	51	51	51	51	51	51
$\overline{R}^2$	.7237	.5147	.6136	.5959	.5088	.4245	.5103	.6675
prob $> F$	.0000	.0000	.0000	.0000	.0000	.0000	.4039	.0034

Table D10 Regressions of exposed to higher education from the census on estimates (standard errors)

Variable	ALL	1940	1950	1960	1970	1980	1990	2000
Weighted								
E	0.9567	0.5182	0.6794	0.8520	0.8963	1.068	1.122	1.121
	(0.008)	(0.061)	(0.062)	(0.069)	(0.057)	(0.049)	(0.088)	(0.042)
Constant	0.030	0.083	0.075	0.032	0.044	−0.015	−0.028	−0.071
	(0.003)	(0.006)	(0.009)	(0.014)	(0.013)	(0.017)	(0.041)	(0.023)
N	355	49	51	51	51	51	51	51
ρ	.9883	7284	.7890	.8123	.9034	.9354	.9183	.9533
prob $> F$	.0000	.0000	.0000	.0758	.0000	.0000	.0000	.0046
Unweighted								
E	0.9354	0.6440	0.7931	0.8632	0.9850	1.076	1.059	1.054
	(0.008)	(0.087)	(0.087)	(0.087)	(0.066)	(0.057)	(0.064)	(0.047)
Constant	(0.038)	0.076	0.062	0.039	0.026	−0.014	−0.011	−0.031
	(0.003)	(0.009)	(0.013)	(0.017)	(0.016)	(0.020)	(0.031)	(0.026)
N	355	49	51	51	51	51	51	51
$\overline{R}^2$	.9767	.5306	.6225	.6598	.8161	.8750	.8432	.9087
prob $> F$	.0000	.0000	.0000	.0007	.0000	.0004	.0001	.0034

These late starters and repeaters may also help to explain our slightly lower secondary exposure rates. Students who start late or repeat elementary grades may be more likely to attend school only until they are legally required. A late starter or repeater who is legally require to attend until 16 years of age may drop out of secondary school at a lower grade than students who have started on time and progressed without repeating. Our use of δ_t accounts for the higher attrition rates, but we still slightly understate secondary exposure. Our weighted and unweighted means are always within 0.03, which is less than a 6% deviation from the census mean.

Finally, our fit of higher education shares is excellent with the exception of 1940 and to a much lesser degree 1950. From 1960 onward, as higher education begins to play a larger roll in the overall level of schooling, we are close to the means and standard deviations reported by the Census.

An alternative way to compare our estimates with the Census data is to regress each exposure category on the Census data. Tables D8–D10 present our regressions of the census shares on our calculated shares pooled and for each decade from 1940 to 2000 respectively. For elementary exposure, the non-pooled correlations exceed 0.85 twice and exceed 0.70 four other times. For secondary exposure, the correlations never exceed 0.85, but they exceed 0.75 twice and exceed 0.70 four more times. For higher education exposure, all seven cross section correlations exceed 0.7 with four cases exceed 0.9. In none of these cases do our correlations fall below 0.6. The pooled correlations exceed 0.85 for each education category.

References

Angrist, J. D., & Krueger, A. B. (1991). Does compulsory school attendance affect schooling and earnings? *Quarterly Journal of Economics, 106*, 979–1014.

Baier, S., Dwyer, G., & Tamura, R. (2006). How important are capital and total factor productivity for economic growth? *Economic Inquiry*.

Barro, R., & Lee, J.-W. (1993). International comparisons of educational attainment. *Journal of Monetary Economics, 32*, 363–394.

Berry, W. D., Fording, R. C., & Hanson, R. L. (2000). An annual cost of living index for the American States, 1960–95. *Journal of Politics, 60*, 550–567.

Blundell, R., & Bond, S. R. (1998). Initial conditions and moment restrictions in dynamic panel data models. *Journal of Econometrics, 87*, 115–143.

Blundell, R., & Bond, S. R. (1999). GMM estimation with persistent panel data: An application to production functions. The Institute for Fiscal Studies. Working Paper W99/4.

Card, D. (1995). Using geographic variation in college proximity to estimate the return to schooling. In L. N. Christofides, E. K. Grant, & R. Swidinsky (Eds.), *Aspects of labour market behaviour: Essays in honour of John Vanderkamp*, (pp. 201–222). Toronto, Canada: University of Toronto Press.

Denison, E. (1962). *The sources of economic growth in the United States and the alternatives before US*. New York: Committee for Economic Development.

Easterlin, R. (1960a). Regional growth of income: Long term tendencies, 1880–1950. In S. Kuznets, A. R. Miller, & R. Easterlin (Eds.), *Population redistribution and economic growth, United States, 1870–1950*, Vol. 1. *Analyses of Economic Change*, Philadelphia: American Philosophical Society.

Easterlin, R. (1960b). Interregional differences in per capita income, population, and total income, 1840–1950. In W. N. Parker (Ed.), *Trends in the American economy in the Nineteenth century*, Princeton: Princeton University Press.

Easterly, W., & Levine, R. (2001). It's not factor accumulation: Stylized facts and growth models. *World Bank Economics Review, 15*, 177–219.

Fishlow, A. (1966). The common school revival: Fact or fancy? In H. Rosovsky, (Ed.), *Industrialization in two systems* (Essays in honor of Alexander Gerschenkron), (pp. 40–67). John Wiley & Sons.

Goldin, C. (1999). America's graduation from high school: The evolution and spread of secondary schooling in the twentieth century. *Journal of Economic History, 58*, 345–374.

Goldin, C., & Katz, L. (2000). Education and income in the early 20th century: Evidence from the prairies. *Journal of Economic History, 60*, 782–818.

Goldin, C., & Margo, R. A. (1992). Wages, prices, and labor markets before the civil war. In C. Goldin, & H. Rockoff (Eds.), *Strategic factors in nineteenth century American economic history: A volume to honor Robert W. Fogel*, Chicago: NBER, University of Chicago Press.

Gordon, R. (1999). *Macroeconomics*. New York: Addison-Wesley.

Klenow, P. J., & Rodriguez-Clare, A. (1997). The neoclassical revival in growth economics: Has it gone too far? *NBER Macroeconomics Annual*, 73–114.

Kuznets, S. (1946). *National income: A summary of findings*. New York: National Bureau of Economic Research.

Long, C. D. (1958). *The labor force under changing income and employment*. Princeton: Princeton University Press.

Mitchener, K. J., & McLean, I. W. (1997). U.S. regional growth and convergence 1880–1980. *Journal of Economic History, 59*, 1016–1042.

Mulligan, C., & Sala-i-Martin, X. (1997). A labor-income based measure of the aggregate value of human capital. *Journal of Japan and the World Economy, 9*, 159–191.

Mulligan, C., & Sala-i-Martin, X. (2000). Measuring aggregate human capital. *Journal of Economic Growth, 5*, 215–252.

National Catholic Education Association. (various years). *United States Catholic elementary and secondary schools*, Washington, DC: National Catholic Education Association.

Schultz, T. (1961). Education and education growth. In N. B. Henry (Ed.), *Social forces influencing American education*. Chicago: University of Chicago Press.

Snyder, T., Hoffman, L., & Geddes, C. (1998). *State comparisons of education statistics: 1969–70 to 1996–97*, Washington, D.C.: National Center for Education Statistics: U.S. Department of Education.

Staiger, D., & Stock, J. H. (1997). Instrumental variables regression with weak instruments. *Econometrica, 65*, 557–586.

Tamura, R. (2001). Teachers, growth and convergence. *Journal of Political Economy, 109*, 1021–1059.

Temple, J. (1999). The new growth evidence. *Journal of Economic Literature, 37*, 112–156.

Towne, M. W., & Rasmussen, W. D. (1960). Farm gross product and gross investment in the nineteenth century. In *Trends in the American economy in the nineteenth century studies in income and wealth*, (Vol. 24). Princeton, NJ: Princeton University Press.

Turner, C., Tamura, R., Mulholland, S., & Baier, S. (2006). How important are physical capital, human capital and total factor productivity for economic growth? Clemson University working paper.

United States Census Bureau (various years). *Statistical abstracts of the United States*. Washington, DC: U.S. Government Printing Office.

United States (1845). Congress. Senate. By American Statistical Association. 28th Cong., 2nd sess. Senate. Document No. 5. Washington: GPO.

United States Department of Commerce (1975). *Historical statistics of the United States: Colonial times to 1970*. Washington, DC: U.S. Government Printing Office.

United States Department of Education (various years). *Digest of education statistics*. Washington, DC: U.S. Government Printing Office.

United States Department of Health, Education and Welfare. (various years). *Projections of educational statistics to ...*. Washington, DC: U.S. Government Printing Office

Weiss, T. (1999). Estimates of white and nonwhite gainful workers in the United States by age group and sex, 1800 to 1900. *Historical Methods,* 21–36.

Williamson, J. G., & Linder, P. H. (1980). *American inequality*, (pp. 97–132). Academic Press.

Wooldridge, J. (2002). *Econometric analysis of cross section and panel data*. Cambridge, MA: M.I.T. Press.