Congressional Budget Office

Background Paper

Assigning Education Status in CBO's Long-Term Microsimulation Model

October 2008

THE CONGRESS OF THE UNITED STATES

Pub. No. 3141

CBO

Assigning Education Status in CBO's Long-Term Microsimulation Model

October 2008

The Congress of the United States ■ Congressional Budget Office

Notes

Unless otherwise noted, all years are calendar years.

Numbers in the text and tables may not add up to totals because of rounding.

Preface

The Congressional Budget Office (CBO) uses a long-term microsimulation model to analyze the budgetary and distributional effects of Social Security and other age-related federal policies and programs. This background paper provides a detailed description of one important component in CBO's microsimulation model: the method by which CBO imputes educational attainment to the microsimulation population. The text describes the estimated relationships in the model, the data sets used, the principles underlying causal effects, and the properties of the projections. A Web appendix, which provides additional data, is available at www.cbo.gov. In keeping with CBO's mandate to provide objective, nonpartisan analysis, the paper makes no policy recommendations.

Jonathan A. Schwabish wrote the paper under the supervision of Joyce Manchester. Patrick Bernhardt, Ed Harris, Michael Simpson, and Julie Topoleski provided helpful comments. Kate Kelly edited the manuscript, which Maureen Costantino prepared for publication. Allan Keaton prepared the tables and figures with assistance from Carl Mueller. Christine Bogusz proofread the paper. Lenny Skutnik produced the printed copies, Linda Schimmel coordinated the print distribution, and Simone Thomas prepared the electronic versions for www.cbo.gov.

Peter R. Orszag
Director

October 2008

Contents

Figures

Figures (Continued)

Tables

Assigning Education Status in CBO's Long-Term Microsimulation Model

Introduction

This background paper describes the methods the Congressional Budget Office (CBO) uses to assign educational attainment in its long-term microsimulation model (known as the CBOLT model). Education is a key component of several demographic and economic attributes used in the model; among the others are fertility, mortality, labor force participation, and earnings. CBO developed a microsimulation approach for analyzing Social Security and other long-term policy issues in order to provide the Congress with comprehensive analyses of the budgetary, distributional, and aggregate economic aspects of various policy choices. Microsimulation allows analysts to examine how public policy affects Social Security finances under current law and to project what would occur under various proposed alternatives.

The methodological strategy of microsimulation is to generate realistic demographic and economic outcomes for a representative sample of the population. The simulation then applies Social Security tax and benefit rules as a method for identifying the likely effects of current law and of various policy alternatives. Educational attainment is the first external demographic characteristic assigned to the individual and thus is important to help capture the appropriate relationships when marital status, fertility, and labor force participation, for example, are assigned. The projected labor market outcomes in turn determine earnings levels and earnings growth and, ultimately, payroll taxes paid and benefits received by individual workers. Because projected individual earnings determine aggregate revenues and outlays, they are a crucial component in the distribution of taxes and benefits for the entire population.

The core data file used in the CBOLT model contains information on individuals' earnings, Social Security benefit status, age, and sex. Thus, any imputation and assignment of educational attainment must, at most, be based exclusively on those four variables. The advantage of randomly assigning education status to the core data file is that the educational attainment distribution then exactly matches the distribution of educational attainment observed in the population. Such random assignment, however, also might generate an inaccurate distribution of earnings across (and within) different categories of educational attainment. This background paper describes the methods for imputing educational attainment for the representative sample in the CBOLT model. The methods are designed to generate distributions of education that are close to those observed in survey data and to accurately capture the relationship between educational attainment and earnings.

Development of the Model

The CBOLT microsimulation model starts with data from a representative sample of the population and projects demographic and economic outcomes for that sample through time. The root sample is drawn from administrative data under the framework used by the Social Security Administration to create the Continuous Work History Sample (CWHS), which contains the core records of administrative earnings used in the CBOLT model. The Social Security Administration furnishes several data sets to CBO for use in the microsimulation model.[1]

Basic demographic assignments in the CBOLT model include educational attainment, marital transitions and partner assignments, fertility, disability, and eventual death. The economic processes in the model include a series of labor force and earnings modules (labor force participation, full-time or part-time employment, hours worked, periods of unemployment, and earnings) and subsequent tax and benefit calculations. Educational attainment is assigned according to the methodology outlined below for individuals for whom actual earnings data exist in the CWHS; for individuals who enter the model during the projection period, educational attainment is assigned randomly according to frequencies observed in survey data.

Each of the two processes for assigning educational attainment to the CBOLT sample is determined by a person's age in 2004, the last year of the CWHS currently used. For people born after 1939 and before 1977 (who were therefore 28 to 64 years old in 2004), the stream of earnings in the CWHS is compared with a representative stream of earnings predicted by a model that uses 30 years of annual data from the March Current Population Survey (CPS). The stream of earnings over a worker's lifetime—the worker's age–earnings profile—typically rises at younger ages, reaches a plateau during middle age, and then declines as a worker nears and enters retirement. That pattern, which is in the shape of an inverted "U," represents the average, and it differs by educational attainment and by sex. Deviations from the pattern are not unexpected and, therefore, various adjustments are made to the initial assignments in the microsimulation model.

The second process assigns educational attainment to people who are younger than 28 or older than 64. Because the younger group has just a few years of actual earnings, educational attainment is assigned randomly. The earnings patterns for people over the age of 64 appear to differ by education categories in ways that are not demonstrated

1. The other administrative data used to construct the CBOLT root data file include the Detailed Earnings Record, the Summary Earnings Record, the Numident (or the Social Security Numerical Identification System), and the Master Beneficiary Record. For this analysis, the root data file is called the CWHS. Panis and colleagues (2000) discuss the administrative data files in more detail. For more information about the CBOLT root data set (see Appendix B, CBO [2006]).

among later birth cohorts. The two groups are randomly assigned an education status using 11 years of pooled data from the CPS. Similarly, for people who were "born" in the model—the individuals created after the CWHS ends in 2004—educational attainment is randomly assigned at birth to mirror the distribution observed in the general population.

Educational attainment for the randomly assigned groups is estimated by sex and nativity (that is, whether the place of birth is the United States or some other country), using pooled CPS data from calendar years 1993 to 2004 (see Table 1). It is worth noting that, starting with the 1975 birth cohort, the shares are fixed; the CBOLT model does not project changes in the education distribution. There are, of course, several trends in the U.S. education system that could significantly affect the proportions of each group in the future. Those factors include the rising cost of postsecondary education (Schwartz and Scafidi 2004); differential high school dropout rates between men and women, whites and minorities, and urban and rural schools (Swanson 2008); and the challenges state and local governments face in spending for education.

Examining Age–Earnings Profiles

People enter the workforce expecting their earnings to increase over most of their lifetime and then to decline just ahead of, and at, retirement. This inverted-U profile for age and earnings is not universal, however, because of changes in family status, disability status, or simple year-to-year variation in earnings. For people with less education, moreover, the inverted-U does not seem to occur; those workers have lower and flatter earnings profiles over a lifetime than do workers with more education. The different patterns generate the appropriate variability in educational attainment and earnings that is used to impute education in the microsimulation model.

Estimated age–earnings profiles are derived on the basis of 30 years' worth of data from the March CPS, estimated separately by sex with separate sample weights for natives and the foreign born. The CPS collects demographic and economic information from about 50,000 households—for 150,000 to 200,000 individuals—each year. The CPS sample used to estimate the age–earnings profiles contains data from 1975 through 2005 (CPS data from 1976 to 2006). Total earnings are defined as the sum of wage and salary earnings, self-employment earnings, and farm earnings. An average wage adjustment then indexes total earnings to a common base year. The average wage index—which combines changes in wages and prices—shows how standards of living differ from one cohort to the next. Simply indexing by price inflation would leave the effect of productivity increases in the data; the goal is to isolate changes associated with age, sex, education, and birth cohort. Overall earnings growth in the estimated age–earnings profiles may therefore appear slower than reported elsewhere in the literature (Beaudry and Green 2000, Murphy and Welch 1990). (The construction of the pooled CPS sample and the average wage index are discussed in Appendix A.)

Individuals are grouped into five birth cohorts: before 1940 ("pre-1940s"), 1940–1949, 1950–1959, 1960–1969, and 1970 and after ("1970s"). Educational attainment is divided into four categories: less than high school, high school graduate, some college, and college graduate; respondents with missing education or with total earnings less than or equal to zero are dropped from the sample.[2] Because nativity status was not recorded in the CPS before 1993, new sample weights are created using the 1970, 1980, 1990, and 2000 census files and the 2004 American Community Survey (provided through the Integrated Public Use Microdata Series). The new sample weights are used to adjust the CPS data before 1993 so that total population counts of natives and foreign-born people approximate those found in the census (the construction of sample weights is explained in Appendix B).

Tracking Raw Earnings Profiles

The pooled CPS data are used to illustrate age–earnings profiles for men and women separately by four education categories and five birth cohorts. The profiles show three main trends: First, as men's educational attainment rises, the differences among cohorts decline—that is, the returns on investment in education are more or less constant from one cohort to the next. For women, however, as educational attainment increases, the differences between cohorts rise as well, suggesting that the returns have increased over time. Second, earnings growth is positively correlated with educational attainment. Third, the age–earnings profiles for women generally are flatter than they are for men.

Raw Age–Earnings Profiles: Men

When real (average wage-adjusted) average earnings for men in each educational attainment group are graphed for each age cohort, two facts are immediately evident (see Figure 1): First, where people have more education, the differences from one earnings profile to the next are smaller. For example, for 40-year-old high school graduates there is a difference of about $7,000 in earnings between the pre-1940s and the 1960s cohorts; earnings for 40-year-old college graduates in the same two cohorts, however, are about the same. Compositional changes within education groups across cohorts are at least partly responsible for the disparities. That is, the portion of native-born male high school dropouts has declined dramatically, from about 35 percent for those born before 1920 to 10 percent for those born in the 1960s. In the same period, the portion of college graduates has approximately doubled, from about 15 percent to about 25 percent of the native-born male population.

2. Educational attainment is based on the highest recorded level of education. High school graduates include those who have a high school diploma or equivalent. People with some college are those with some college but no degree or those with an associate's degree. See Appendix A for more details on the construction of the pooled CPS data set. The education status variable is also recoded for younger survey respondents. For respondents younger than 19 who report some college, education is recoded to high school graduate; for those under age 22 who are college graduates, education is recoded to some college.

The second observation from men's raw age–earnings profiles is that individuals with more education (especially college graduates) experience faster earnings growth as they age than do workers with less education. This is perhaps most clearly seen in the calculation of the growth from the minimum to the maximum earnings for groups with different educational attainment. Take, for example, the 1950s birth cohort, with average earnings for ages 22 to 54. The differences from the minimum to the maximum earnings for each education category for that cohort are 46 percent (less than high school), 58 percent (high school graduate), 184 percent (some college), and 482 percent (college graduate). Thus, although workers with less education might, at age 22, have higher earnings than those with more education, over a lifetime their earnings profile will be flatter.

Over time, there appear to have been opposing shifts in the profiles from older to more recent cohorts. Younger workers with at most a high school diploma earn less than all of the preceding cohorts. For 25-year-olds who finished high school, for example, workers in the most recent cohort earn about 18 percent less than their counterparts did in the 1950s. At the upper end of the education distribution, however, there appears to be a smaller decline in the earnings profiles for younger workers. That change could reflect the aging of the population or it could be attributable to changes in technology benefiting younger, more technically skilled workers (Juhn, Murphy, and Pierce 1993).

Raw Age–Earnings Profiles: Women

The same set of age–earnings profiles for women give slightly different results (see Figure 2). All of the earnings profiles for women are much flatter than they are for men, and younger women with more education appear to experience faster earnings growth than do most of their older counterparts. There is almost no curvature in the age–earnings profile for women with less than a high school education; earnings lie between about $6,000 and $13,000 (compared with the similar group of men, whose earnings are between $10,000 and $27,000). Women who graduate from high school have higher real average earnings—$3,000 to $6,000 more at each age—with more curvature in their profiles than shown for high school dropouts. There tends to be early growth in earnings among women with at least some college education, but only for women who have graduated college does that growth translate to significant gains relative to the other groups.

The differences in the age–earnings profiles among women in the various educational-attainment groups are more pronounced than they are for men. For women who did not complete high school, earnings are relatively flat and approximately the same for all five birth cohorts. For women who have graduated from college, however, there is a significant upward shift in earnings: For 40-year-old women, average earnings grew by more than 14 percent between the 1940s and 1960s birth cohorts. For men, average earnings grew by less than 1 percent across the same groups.

The changes in women's earnings profiles reflect changes in the labor force, perhaps the most important of which is attributable to the sharp increase in women's labor force participation rates over the past 30 years. In 1975, about 44 percent of women were in the labor force; by 2005 that portion neared 60 percent. Men's participation, meanwhile, has fallen somewhat, from about 80 percent to about 75 percent over the same period (Bureau of Labor Statistics 2008). Other changes in the labor force and in society in general—such as patterns of marriage and childbearing, discrimination, labor market competition, and social mores—have resulted in changing roles for women in the labor force and hence in their earnings.

Predicting Age–Earnings Profiles in a Regression Framework

Age–earnings profiles for each education level, cohort, age, and nativity group are predicted by a simple regression model. Using ordinary least-squares regressions, total earnings (*Y*) are predicted using a set of age, education, and cohort dummy variables (*D*), and a full set of dummy variable interactions:[3]

$$Y = \sum_{i=16}^{70} D_{age_i} + \sum_{i=16}^{70} \left(D_{age_i} \times \left(\sum_{t=1}^{4} D_{education_t} \right) \right) + \sum_{i=1}^{5} D_{cohort_i} + \sum_{i=1}^{5} \left(D_{cohort_i} \times \left(\sum_{t=1}^{4} D_{education_t} \right) \right) \quad (1)$$

The regressions are estimated separately for men and women using different weights for natives and the foreign born. Those who are 70 and older and high school graduates constitute the reference group. The R^2 statistic is about 72 percent for each of the four regressions, signaling that the regression accounts for nearly three-quarters of the variation in earnings. (Regression coefficients for all four regressions are presented in the Web Appendix Tables W1 through W4.)

Predicted Age–Earnings Profiles: Men

Using the regression model specified above, predicted earnings for each education and cohort combination are obtained for each age, 16 to 70. Thus, for any cohort for

3. Alternative regressions add the national unemployment rate, which serves to shift the profiles upward but otherwise has little effect on the estimated age–earnings profiles. Using the logarithm of earnings as the dependent variable has little effect on the overall fit of the model. If survey respondents with zero reported earnings are included in the regressions, the profiles demonstrate much larger declines among older groups, for which zero earnings are more likely. For purposes of the regressions, the top and bottom 1 percent of earners were eliminated from the sample.

which there are no raw data, the regression obtains estimates using observations from the other cohorts.[4]

Differences in the age–earnings profile by educational attainment are evident when birth cohort is held constant (see Figure 3 and Web Appendix Figures W1 through W3). For men in the 1940s cohort (Figure 3, top), there is significantly more curvature for college graduates than there is for men with less education. College graduates earn less than other groups from about age 22 to age 26, but the steeper slope of the profile then results in higher earnings for that group for the rest of their work lives. For the 1970s cohort (Figure 3, bottom), the profiles are similar, but the starting points are slightly lower than for the 1940s cohort.[5]

The figures show differences between natives and the foreign born that are generally larger for groups with more education and for more recent cohorts. For the 1940s cohort, for example, the average difference between earnings predicted for natives and foreign-born people between the ages of 30 and 50 was $1,137 (less than high school), $973 (high school graduates), $1,227 (some college), and $2,338 (college graduates). The same metric for the 1970s cohorts yields larger differences for the three highest education groups: $661 (less than high school), $1,079 (high school graduates), $1,519 (some college), and $3,622 (college graduates).[6]

The differences in the predicted age–earnings profiles from one education group to another are well established in the economics literature and reflect increasing returns on the investment in education, changing demands for labor, changes in the return on investment in different sets of skills, and changes in technology (see, for example, Card [1999] and Katz and Murphy [1992]). The estimates visibly demonstrate not

---

4. The profiles are roughly similar to those seen in Beaudry and Green (2000) and in Murphy and Welch (1990), although those studies use log earnings as the dependent variable. Murphy and Welch used years of experience rather than age on the right-hand side of the equation. Because measures of experience must be inferred from an equation (the assumed start of the person's working lifetime, typically the person's age minus 6 minus the number of years of education) and because experience is nearly perfectly correlated with age, the regressions above simply use age as a right-hand-side regressor. Using administrative earnings records, Bosworth, Burtless, and Steuerle (1999) track age–earnings profiles for men and women born between 1931 and 1960. They show that few workers have level career earnings, and they find differences across sex and education groups that are similar to those reported here.

5. A simple regression of the predicted earnings values on age and age-squared shows those differences quantitatively. The coefficient on the age-squared term for each education level for native men is −20.3 (less than high school), −28.3 (high school graduate), −38.9 (some college), and −58.7 (college graduate). For foreign-born men, the analogous coefficients are −19.5, −27.7, −38.1, and −56.6.

6. The regression model does not illustrate some of the important characteristics of the foreign-born population—separately from native-born people—that help determine the path of their age–earnings profiles. Among those characteristics are the person's age in the year of immigration, his or her country of origin, and the eventual place of residence (see, for example, Borjas [1987], Lubotsky [2007], and Passel and Zimmerman [2001]).

only the shift across the education categories but also the disparity in earnings growth for higher and lower educational attainment. Men with a high school degree or less experience virtually no earnings growth after about age 35, and their earnings growth before age 35 is only a fraction of that exhibited among men with at least some college.

Predicted Age–Earnings Profiles: Women

The regression model is estimated separately for the approximately 1 million women in the sample. (Figure 4 shows the estimates for the 1940s [top] and 1970s [bottom] cohorts graphically; Web Appendix Figures W4 through W6 show predicted profiles for the other cohorts.) Female college graduates in the 1940s cohort exhibit annual earnings that exceed $20,000 for most of their working lifetimes. Women in the other three education categories experience similar age–earnings profiles; on average, the difference from one educational group to the next is about $6,000 to $7,000. In the most recent cohort, college graduates experience much faster earnings growth, and the gaps between the other three education levels are only slightly larger than they are for the 1940s cohort.

The age–earnings profiles are flatter for women than they are for men, although earnings for women continue to grow until the mid-50s.[7] Men's profiles, in contrast, grow more quickly up through the mid-30s before flattening. The earnings gap between men and women remains substantial regardless of educational attainment, but it is largest for college graduates. The male–female earnings ratio for 40-year-old high school graduates born in 1970 or later is about 1.6 (see Figures 3 and 4); for college graduates it is about 1.8.

Imputing Education Status in the CBOLT Model

To project *earnings* in the CBOLT model for people whose educational attainment is known, the coefficients from regressions similar to those specified in Equation (1) are imputed to each individual in the microsimulation. Using the individual's age, birth cohort, and educational attainment, initial earnings are then calculated (growth in earnings is calculated in a separate process; see CBO [2006]). To impute *education,* however, the regressors from Equation (1) are reversed so that education becomes an output, with actual earnings, age, and birth cohort as inputs.

The absolute value of the difference between actual earnings in the CWHS and predicted earnings for each education level from the CPS is calculated for each individual in the CBOLT model at each age. Then, at each age, the education level that

7. This is reflected in the simple regression of the predicted earnings values on age and age-squared (see also Footnote 5). The coefficient on the age-squared term for each education level for native women is −9.1 (less than high school), −11.7 (high school graduate), −15.2 (some college), and −24.2 (college graduate). For foreign-born women, the analogous coefficients are −9.1, −11.8, −15.5, and −24.3.

minimizes the absolute value of the difference between actual and predicted earnings is assigned:

$$\text{minimum gap}_{age} = \text{minimum}_{age}\,(|Y - \hat{Y}_{LSHS}|;|Y - \hat{Y}_{HSGD}|;|Y - \hat{Y}_{SOCL}|;|Y - \hat{Y}_{COGD}|) \quad (2)$$

Educational attainment (LSHS is less than high school, HSGD is high school graduate, SOCL is some college, and COGD is college graduate) is determined for people who have positive earnings at those ages and who are not receiving disability insurance benefits. The calculation is done by comparing the probability of each education level to a random number selected from a uniform distribution.

An example might help clarify the process: In this case, consider a man born in the United States in 1954 (thus, age 50 in 2004) who has a stream of earnings over his lifetime (see Table 2). The earnings for this imaginary worker from the age of 30 through 39 are shown in the second column of Table 2. In columns 3 through 6, predicted earnings from the CPS model for males born in the United States in the 1950s are shown for each level of educational attainment. The absolute value of the difference between each predicted education level and actual earnings is then calculated; the absolute value of the minimum across those four values is shown in column 7. Education is then assigned using the frequencies of each education level resulting from the minimum of the absolute difference. Thus, the probability of this worker having less than a high school education is 30 percent (3 out of 10), there is a 60 percent probability that he is a high school graduate, a 10 percent probability that he has some college education, and a zero probability that he is a college graduate. Each probability is compared with a number drawn randomly from a uniform distribution to determine this theoretical person's educational attainment.[8]

To be sure, not everyone's earnings path is so simple or smooth. Some high school graduates' earnings resemble those of the typical college graduate; some college graduates work in low-paying jobs for a lifetime. Moreover, year-to-year variability in earnings can create uneven paths in an individual worker's age–earnings profile. For example, CBO has reported that about 40 percent of workers experience a 25 percent or greater change in their year-to-year earnings (CBO 2008). Thus, the variation in earnings both across and within education categories complicates the initial imputation strategy.

8. Alternatively, the most common occurrence of education (the mode) in the series could be used to assign educational attainment. Thus, in the example, the worker would be considered a high school graduate. That methodology, however, requires a secondary adjustment to account for people who have low earnings even though they actually have more education. The differences between the two methods, however, are small.

Adjusting the CPS-Predicted Profiles

Before making secondary adjustments, it is important to note that there are fundamental differences between the CPS and CWHS data sets. First, the CWHS contains more observations at the very bottom of the distribution and, because of top-coding in the CPS, more observations at the very top of the distribution (see Table 3). For example, 3.5 percent of workers in the CWHS had earnings of $1,000 or less in 2004; only 2.1 percent of the workers in the CPS are in that category. The differences decline as earnings rise: 21.0 percent of workers in the CWHS and 20.6 percent of workers in the CPS had earnings of $10,000 or less in 2004. At the top of the distribution, because of top-coding, there are no workers with earnings of at least $300,000 in the CPS. In the CWHS, however, 0.6 percent of workers have at least $300,000 of earnings. To accommodate the differences, the CPS-predicted profiles are adjusted to better capture the age–earnings profiles of individuals in the CWHS. To do so, each CPS-predicted age–earnings profile is multiplied by a fixed factor such that the profiles are shifted up or down to better match the earnings distribution in the CWHS (see Table 4).

To adjust for differences in the CWHS and CPS and for variability in individual earnings patterns, two more adjustments are made to the initial assignments, based separately on sex and nativity. The first compares actual earnings at different points in a worker's lifetime with earnings predicted by the CPS regression model. For people who have positive earnings for ages 28 to 30 and 52 to 54, the ratio of total actual earnings at ages 52 to 54 relative to total actual earnings at ages 28 to 30 ($Y_{52–54}/Y_{28–30}$) is compared with the same ratio from the predicted earnings profiles. If the ratio of actual earnings is larger than that predicted in the CPS, the person's educational attainment is increased to the next level. If the ratio exceeds an even greater threshold than the CPS prediction, educational attainment is increased by two categories. For people who were 51 and younger in 2004 and had at least one year of zero earnings out of the three years between ages 28 to 30 and again between ages 49 to 51, the points of comparison are sequentially moved closer together. Once three consecutive years of positive earnings are found, the ratio calculation is repeated and the same set of thresholds is applied. For people who were 52 and older in 2004 and had at least one year of zero earnings between 28 and 30 and between 52 and 54, the points of comparison are moved forward. Thus, for the denominator, three years of earnings between the ages of 31 and 54 are used; for the numerator, three years of earnings between the ages of 54 and 70 are used.

The thresholds used to compare the actual ratios with the CPS-predicted ratios differ by sex and nativity (see Table 5). The thresholds were chosen such that the education distribution in the CBOLT model approximated that in the CPS. The one-step education adjustment is made for those whose ratios fall between the lower and upper thresholds; the two-step adjustment is made for those whose ratios exceed the upper threshold.

A second adjustment is made to correct for the possibility that some women temporarily leave the labor force for childbearing. Because those women reduce their work time or stop working altogether to raise children and later return to the workforce, their earnings profiles are flatter than might otherwise be expected. Similarly, the ratio of earnings at older ages to earnings at younger ages is lower than expected. To adjust for such earnings patterns, the ratio of average earnings during ages 52 to 54 to average earnings during ages 26 to 28 ($Y_{52–54}/Y_{26–28}$) is calculated for women who have positive earnings for ages 26 to 28 and then again at some point from ages 52 to 54. Half of the women initially assigned an educational attainment of less than high school or high school degree whose earnings ratio is less than one are randomly assigned to the some-college category; the other half are assigned to the college-graduate category.

Imputed Education Status Compared with Observed Survey Data

After educational attainment is assigned to everyone in the root data file, the education distributions are compared with those in the CPS. The goal is to come as close to the actual CPS education distribution as possible so the subsequent distribution of earnings mirrors that found in the survey data. No adjustments are made to the earnings data, however, to close any remaining gaps.

According to the CPS, the percentage of native men who did not finish high school declined swiftly in the first half of the 20th century, from more than 40 percent for men who were born in 1914 to about 9 percent among those born in 1950 (see Figure 5); the percentage who completed high school went largely unchanged over the period. The imputation methodology approximates those education shares for people born between 1940 and 1977, although the line is flatter than that found in the CPS. Recall that people who are older than 64 or younger than 28 in 2004 (born before 1939 or after 1977) are randomly assigned an education status targeted to shares computed from pooled data from the CPS from calendar years 1993 to 2004.[9]

The decline in the share of high school dropouts among cohorts born in the first half of the century is matched by increases in educational attainment in the same groups. About 15 percent of native men born in 1914 had some college education; more than 30 percent of those born in the late-1970s had some college education (see Figure 5). The imputations match that trend fairly well, with no differences of more than 8 percentage points between any two points. Similar increases are seen among those who had completed college—the percentage of workers with a college degree rose from about 15 percent to between 25 percent and 30 percent over the period. Again, the imputations more or less match the trends observed in CPS data by year of birth.

Changes in the distribution of educational attainment for women are similar (see Figure 6). The percentage of native-born women with some college education or a

9. For the younger groups, the CPS education targets are fixed to match those for the 1975 birth cohort.

college degree grew substantially during the 20th century. The share of women with a college degree went from about 10 percent of those born in 1914 to about 33 percent of those born in 1975. Similarly, the share that had some college increased from 15 percent for women born in 1914 to more than 30 percent for those born after 1975. The imputation estimates mirror those trends, with slightly greater errors for the highest education group of women born during the 1970s than of those born during the 1940s.

The share of native-born women who graduated high school grew between the 1914 and 1938 birth cohorts before declining steadily for the next 30 years of births (see Figure 6). About 40 percent of women born in 1940 finished their education as high school graduates; only about a quarter of women born in the mid-1970s had the same level of education. Much faster declines are found in the number of women who never finished high school. That proportion dropped from almost 45 percent for the 1914 birth cohort to less than 10 percent for the most recent birth cohorts. Again, the imputations more or less follow the observed patterns, although slightly flatter patterns are predicted.

Educational attainment for the foreign-born population basically follows the patterns established among natives. However, because there are fewer foreign-born individuals in the CBOLT model's sample, the distribution of education imputed for them is more variable. (Analogous figures for the educational share for foreign-born males and females by year of birth are presented online in the Web Appendix Figures W7 through W14.)

Distribution of Earnings

After educational attainment is imputed for the sample, the next step is to compare the distribution of earnings with that observed in the CPS. An important caveat, however, is that regardless of the educational attainment imputed, the distribution of earnings in the two samples differs in fundamental ways (see Table 3 and Schwabish [2006]). Thus, the comparisons (see Tables 6 through 9) are restricted to earnings above $872, the real-dollar value of the minimum quarterly earnings required to qualify for Social Security benefits. Earnings also are restricted to less than $280,000, the CPS-adjusted earnings top-code level in 2004. (Appendix A gives more details about top-coding. An analogous secondary set of tables, using earnings ranges restricted only from the bottom, is included in Web Appendix Tables W5 through W8. Web Appendix Tables W9 through W12 present the comparisons of restricted and unrestricted median earnings.)

Average Earnings. The difference between average earnings in the CBOLT model and the CPS generally declines as education increases (see Tables 6 and 7). The random-assignment method (for people older than 64 or younger than 28) shows the opposite pattern—the differences between imputed CBOLT and actual CPS earnings are larger for the upper education groups. Those differences, however, have less importance for the overall microsimulation model because most of the lifetime earnings for younger people will be projected using the earnings methodology described in CBO (2006) once education status is assigned. The work histories of older workers randomly assigned to an educational attainment group are almost completely set, and thus benefits are computed using observed, actual earnings.

The difference between overall average earnings in the CBOLT model and the CPS for native-born male college graduates is small, at most 13 percent (see Table 6). For the lower education categories, average earnings are about 30 percent smaller in the CBOLT model's sample, which suggests significantly more earnings at the bottom of the distribution than in the CPS. This is further reflected in the "All Groups" category, which shows lower overall average earnings in the CBOLT model sample. Web Appendix Tables W9 through W12 further illustrate the differences in median earnings that support this claim: In the CBOLT model, median earnings for those who do not finish high school are much lower than are average earnings even though the two categories are much closer in the CPS. This suggests a much larger skew toward workers with lower earnings in the CBOLT model's earnings distribution among this low-education group. There appears to be more clustering at the lower tail in the CBOLT earnings distribution than in the CPS.

For foreign-born men, the differences between average earnings in the CBOLT model and the CPS are generally smaller than they are for native-born men. Average CBOLT earnings for foreign-born high school graduates are smaller than the CPS averages by 26 percent (for 45- to 54-year-olds). The difference between CBOLT and CPS average earnings tends to broaden as the foreign born age, whereas the opposite pattern emerges for native men. For foreign-born men with less than a high school education, for example, average earnings for 28- to 34-year-olds are about 8 percent lower than in the CPS; average earnings for 55- to 64-year-olds are 46 percent lower than in the CPS. For native men, those differences narrow, decreasing from 48 percent to 22 percent. The other three education groups exhibit roughly similar patterns.[10]

The differences in the two sets of average earnings are slightly smaller for foreign-born than for native-born women (see Table 7). Relative to men, women's average earnings are typically closer to the CPS estimates for all groups other than college graduates. One possible reason for this difference is that fewer women earn near or above the $280,000 top-code level. Earnings for college-educated, native-born women are about the same in the CBOLT model and the CPS except for women ages 28 to 34, who earn 21 percent more. For high school graduates, average earnings in the CPS are about 30 percent higher than are those from the CBOLT model. Similar to men, the differences between the two sets of average earnings among foreign-born women expand with age, perhaps because of fundamental differences in actual earnings

10. There is a difference of 10 percent to 15 percent in the average earnings for the entire distribution, regardless of the education status. Thus, for most groups, the difference between average earnings in the CBOLT model and the CPS samples is mitigated by underlying differences in the entire distribution.

profiles among the foreign born (even though the differences in the CPS-predicted profiles do not suggest that this is the case).

Standard Deviation of Earnings. The pattern of earnings dispersion within each education–age group cell differs from that found in average earnings. Among native-born male high school dropouts, the dispersion in earnings is about 35 percent larger in the CBOLT model than in the CPS; again, however, that might reflect different clustering at the bottom of the distribution (see Table 8). For native- and foreign-born men alike, the dispersion in earnings is greatest among those ages 35 to 44. The standard deviation of earnings in the CBOLT model for high school graduates in that age group is about $14,000; the analogous estimate in the CPS is about $18,000, a difference of 24 percent for native men and 19 percent for foreign-born men. The differences in levels of dispersion in the CPS and the CBOLT model between the two nativity groups could occur because age–earnings profiles of the foreign born differ from those of natives in ways that are not properly predicted in the regression model.[11] Alternatively, the differences could arise from the smaller sample of foreign-born people in the CBOLT model.

In general, differences in the dispersion of earnings for women in the CPS and the CBOLT model are just as large as they are for men (see Table 9). Among college graduates, however, the standard deviation of earnings is almost identical in the two data sets (except for foreign-born women ages 35 to 44; that group shows a difference of 24 percent). Similar to the estimates for men, the largest difference in the standard deviation in earnings is found among women with some college education. That larger gap could reflect the greater probability that those in the some-college category are a mix between true high school graduates and true college graduates.[12]

11. Foreign-born people often start their working lives in the United States at substantially lower earnings but experience faster earnings growth than do native-born people with similar years of education and work history (Borjas 1989, Duleep and Dowhan 2008, and Lubotsky 2007). The use of repeated cross-sectional earnings data (as in the CPS) to predict age–earnings profiles of the foreign born (as opposed to longitudinal earnings data, as in the CBOLT model) might not accurately capture those differences.

12. There are large differences in the standard deviation of earnings when unrestricted earnings samples are used (see Web Appendix Tables W7 and W8). For example, the difference in the standard deviation of earnings for 35- to 44-year-old male college graduates in the restricted sample in Table 8 is 18 percent; in the unrestricted sample in Table W7, that figure increases to 348 percent. The larger (and longer) tail in the CBOLT model is most likely responsible for the differences.

Table 1.

Distribution of Educational Attainment for Random Assignment in CBO's Long-Term Microsimulation Model

(Percent)

	Native		Foreign-Born	
	Men	Women	Men	Women
Less Than High School	9.6	8.4	35.0	30.3
High School Graduate	31.3	25.3	25.0	22.0
Some College	30.3	33.1	15.3	16.5
College Graduate	28.9	33.1	24.7	31.2

Sources: Congressional Budget Office based on data from the March Current Population Survey, 1994 to 2005.

Note: Individuals randomly assigned educational attainment are born before 1939 or after 1977.

Table 2.

Sample Educational Imputation Methodology

(1993 Dollars)

		CPS Predictions					
Age	Actual Earnings	Less than High School	High School Graduate	Some College	College Graduate	Minimum Difference (Absolute value)	Resulting Educational Attainment
30	21,000	22,034	29,556	32,201	36,284	1,034	LSHS
31	23,000	21,972	30,361	32,992	38,545	1,028	LSHS
32	25,000	21,866	30,441	33,957	40,817	3,134	LSHS
33	28,000	22,610	30,983	34,430	42,265	2,983	HSGD
34	29,000	22,791	31,335	35,321	43,935	2,335	HSGD
35	30,000	23,236	31,766	36,162	46,576	1,766	HSGD
36	35,000	23,385	32,239	36,765	47,182	1,765	SOCL
37	34,000	22,475	32,335	36,410	48,189	1,665	HSGD
38	33,000	23,280	32,329	37,251	49,528	671	HSGD
39	33,000	23,415	32,457	37,606	49,943	543	HSGD

Sources: Congressional Budget Office's long-term microsimulation model and the March CPS, 1976 to 2005.

Notes: CPS = Current Population Survey; LSHS = less than high school; HSGD = high school graduate; SOCL = some college.

Table 3.

Workers with Positive Earnings Below or Above Earnings Cutoffs in 2004

(Percent)

Earnings (Dollars)	CWHS Men	CWHS Women	CWHS Total	CPS Men	CPS Women	CPS Total
≤ 1,000	3.0	4.1	3.5	1.4	3.0	2.1
≤ 2,000	4.9	7.0	5.9	2.4	5.5	3.8
≤ 3,000	6.6	9.6	8.0	3.7	8.2	5.8
≤ 4,000	8.2	11.9	10.0	4.8	10.7	7.6
≤ 5,000	9.8	14.2	11.8	6.0	13.5	9.5
≤ 10,000	17.4	25.1	21.0	14.2	27.8	20.6
≥ 100,000	7.6	2.1	5.0	3.6	0.7	2.2
≥ 200,000	2.0	0.4	1.3	0.0	0.0	0.1
≥ 300,000	0.9	0.1	0.6	0.0	0.0	0.0

Sources: Congressional Budget Office's long-term microsimulation model and the March 2005 CPS.

Note: CWHS = Continuous Work History Sample; CPS = Current Population Survey.

Table 4.

Multipliers Used to Adjust CPS-Predicted Earnings Profiles, by Nativity Group

	Native		Foreign Born	
	Men	Women	Men	Women
Less Than High School	0.10	0.10	0.30	0.30
High School Graduate	0.35	0.10	0.90	0.60
Some College	1.10	1.25	1.00	1.00
College Graduate	1.00	1.00	0.90	1.00

Source: Congressional Budget Office.

Note: CPS = Current Population Survey.

Table 5.

Thresholds to Adjust the Ratio of Earnings

	Native		Foreign Born	
	Men	Women	Men	Women
Lower Threshold	3.0	3.0	1.5	1.5
Upper Threshold	5.0	5.0	2.0	2.0

Source: Congressional Budget Office.

Table 6.

Average Earnings Comparison for Men, 2004

(Dollars)

Age	Native: Less Than High School	Native: High School Graduate	Native: Some College	Native: College Graduate	Native: All Groups	Foreign Born: Less Than High School	Foreign Born: High School Graduate	Foreign Born: Some College	Foreign Born: College Graduate	Foreign Born: All Groups
					Long-Term Microsimulation Sample					
22 to 27[a]	14,440	14,608	14,794	14,758	14,691	14,531	15,972	15,334	14,499	15,021
28 to 34	9,410	14,743	23,936	44,009	25,249	12,287	18,591	25,358	46,716	24,587
35 to 44	13,306	18,395	28,929	55,296	31,870	14,522	21,290	29,036	53,459	26,721
45 to 54	13,758	19,398	33,968	55,782	35,363	10,976	17,796	27,241	54,014	27,578
55 to 64	14,481	19,404	31,540	48,159	33,328	9,843	16,330	24,344	49,433	26,569
65+[a]	15,997	19,107	16,745	17,301	17,455	22,337	19,004	28,813	18,190	21,143
					Current Population Survey					
22 to 27[a]	14,238	17,025	16,201	23,058	17,804	12,757	13,759	14,544	19,882	14,339
28 to 34	17,963	22,528	27,059	38,980	29,011	13,412	17,380	22,805	38,087	22,313
35 to 44	19,141	26,881	33,765	55,896	37,580	17,374	21,867	30,616	49,548	29,739
45 to 54	19,618	28,415	35,286	58,603	39,839	16,255	23,961	28,843	49,521	30,728
55 to 64	18,634	26,864	33,865	52,218	37,640	18,319	22,114	29,750	46,925	30,434
65+[a]	14,357	15,908	22,319	35,206	23,613	13,439	19,387	19,187	43,283	26,428
					Difference (Percent)					
22 to 27[a]	1	-14	-9	-36	-17	14	16	5	-27	5
28 to 34	-48	-35	-12	13	-13	-8	7	11	23	10
35 to 44	-30	-32	-14	-1	-15	-16	-3	-5	8	-10
45 to 54	-30	-32	-4	-5	-11	-32	-26	-6	9	-10
55 to 64	-22	-28	-7	-8	-11	-46	-26	-18	5	-13
65+[a]	11	20	-25	-51	-26	66	-2	50	-58	-20

Source: Congressional Budget Office based on data from the March 2005 Current Population Survey and from CBO's long-term microsimulation model.

Note: Samples restricted to individuals with earnings above $872 and below $280,000 in 2004. Imputed educational attainment is derived from incorporating estimated regression results from pooled March Current Population Survey files into CBO's long-term microsimulation model. See the text for details.

a. Education status is assigned randomly using pooled Current Population Survey data from 1994 to 2005.

Table 7.

Average Earnings Comparison for Women, 2004

(Dollars)

Age	Native: Less Than High School	Native: High School Graduate	Native: Some College	Native: College Graduate	Native: All Groups	Foreign Born: Less Than High School	Foreign Born: High School Graduate	Foreign Born: Some College	Foreign Born: College Graduate	Foreign Born: All Groups
					Long-Term Microsimulation Sample					
22 to 27[a]	11,712	11,965	12,301	12,694	12,299	12,166	12,239	13,486	12,497	12,506
28 to 34	6,811	9,378	17,069	33,115	18,740	8,110	11,447	17,000	34,081	18,176
35 to 44	8,039	11,371	19,311	37,033	21,115	8,805	12,505	18,785	38,605	19,883
45 to 54	10,808	12,476	21,047	36,194	22,991	7,394	11,408	18,652	35,045	20,129
55 to 64	11,048	11,790	18,231	29,985	20,451	6,496	9,860	16,807	31,127	19,680
65+[a]	10,608	11,394	11,248	11,520	11,216	12,315	12,488	18,052	14,721	13,490
					Current Population Survey					
22 to 27[a]	8,655	11,990	12,606	18,775	14,064	8,238	10,629	13,844	19,420	12,922
28 to 34	10,516	15,211	17,128	27,278	20,456	9,515	12,516	18,076	25,402	17,502
35 to 44	12,394	16,566	20,272	32,356	22,852	10,026	15,373	18,721	33,639	20,395
45 to 54	12,755	18,203	22,583	34,650	24,767	11,253	15,116	21,815	30,126	20,106
55 to 64	11,179	16,488	20,919	32,585	22,462	10,494	16,804	23,640	30,036	21,021
65+[a]	7,396	11,853	14,651	19,547	13,527	10,313	14,066	10,349	20,679	15,066
					Difference (Percent)					
22 to 27[a]	35	0	-2	-32	-13	48	15	-3	-36	-3
28 to 34	-35	-38	0	21	-8	-15	-9	-6	34	4
35 to 44	-35	-31	-5	14	-8	-12	-19	0	15	-3
45 to 54	-15	-31	-7	4	-7	-34	-25	-14	16	0
55 to 64	-1	-28	-13	-8	-9	-38	-41	-29	4	-6
65+[a]	43	-4	-23	-41	-17	19	-11	74	-29	-10

Source: Congressional Budget Office based on data from the March 2005 Current Population Survey and from CBO's long-term microsimulation model.

Note: Samples restricted to individuals with earnings above $872 and below $280,000 in 2004. Imputed educational attainment is derived from incorporating estimated regression results from pooled March Current Population Survey files into CBO's long-term microsimulation model. See the text for details.

a. Education status is assigned randomly using pooled Current Population Survey data from 1994 to 2005.

Table 8.

Standard Deviation of Earnings Comparison for Men, 2004

(Dollars)

	Native					Foreign Born				
Age	Less Than High School	High School Graduate	Some College	College Graduate	All Groups	Less Than High School	High School Graduate	Some College	College Graduate	All Groups
	Long-Term Microsimulation Sample									
22 to 27[a]	10,644	11,264	12,033	11,383	11,480	11,362	13,561	12,053	15,020	13,057
28 to 34	11,969	10,579	11,942	23,902	20,235	11,824	11,903	9,435	24,226	20,734
35 to 44	16,862	13,980	15,095	31,831	26,305	15,495	14,356	13,728	30,534	23,907
45 to 54	15,235	14,345	20,662	35,037	28,955	10,147	13,819	17,260	36,542	27,594
55 to 64	17,508	15,042	23,617	37,879	30,413	7,249	12,826	18,843	38,933	28,697
65+[a]	23,459	27,646	23,597	23,459	24,948	30,837	23,860	34,462	20,597	27,411
	Current Population Survey									
22 to 27[a]	12,443	10,985	11,723	16,839	13,162	7,310	12,654	9,193	18,741	11,616
28 to 34	12,013	13,802	17,197	25,533	20,609	7,083	9,412	16,844	24,061	18,258
35 to 44	12,842	18,276	22,716	38,733	30,325	11,905	17,778	25,785	36,937	28,115
45 to 54	12,781	18,836	24,647	41,193	32,402	9,718	21,242	21,118	38,819	29,645
55 to 64	12,701	19,561	25,086	41,206	32,766	17,508	13,513	21,268	38,798	28,908
65+[a]	17,952	16,418	26,236	40,842	30,270	7,824	15,907	11,113	47,435	32,871
	Difference (Percent)									
22 to 27[a]	-14	3	3	-32	-13	55	7	31	-20	12
28 to 34	0	-23	-31	-6	-2	67	26	-44	1	14
35 to 44	31	-24	-34	-18	-13	30	-19	-47	-17	-15
45 to 54	19	-24	-16	-15	-11	4	-35	-18	-6	-7
55 to 64	38	-23	-6	-8	-7	-59	-5	-11	0	-1
65+[a]	31	68	-10	-43	-18	294	50	210	-57	-17

Source: Congressional Budget Office based on data from the March 2005 Current Population Survey and from CBO's long-term microsimulation model.

Note: Samples restricted to individuals with earnings above $872 and below $280,000 in 2004. Imputed educational attainment is derived from incorporating estimated regression results from pooled March Current Population Survey files into CBO's long-term microsimulation model. See the text for details.

a. Education status is assigned randomly using pooled Current Population Survey data from 1994 to 2005.

Table 9.

Standard Deviation of Earnings Comparison for Women, 2004

(Dollars)

	Native					Foreign Born				
Age	Less Than High School	High School Graduate	Some College	College Graduate	All Groups	Less Than High School	High School Graduate	Some College	College Graduate	All Groups
					Long-Term Microsimulation Sample					
22 to 27[a]	8,911	9,163	9,303	10,268	9,576	9,546	9,947	12,363	11,171	10,665
28 to 34	9,431	9,332	8,983	18,486	15,715	9,467	8,303	9,376	17,499	15,804
35 to 44	8,526	9,967	11,132	23,694	18,591	10,458	10,574	11,981	23,527	19,349
45 to 54	11,695	10,487	13,055	24,167	19,501	6,711	7,459	14,003	23,429	19,349
55 to 64	13,463	9,822	12,195	24,168	18,522	7,871	6,366	12,502	25,221	20,359
65+[a]	14,437	14,000	14,045	16,426	14,492	15,534	13,573	28,643	13,436	17,188
					Current Population Survey					
22 to 27[a]	6,638	8,376	8,259	13,479	10,569	5,536	6,618	13,781	14,127	11,536
28 to 34	7,654	13,184	11,602	18,673	16,070	6,110	8,443	13,536	18,955	15,192
35 to 44	10,872	12,238	13,529	24,059	18,696	6,588	11,932	16,124	30,757	21,812
45 to 54	11,900	13,516	15,302	24,772	19,781	9,357	9,685	17,690	23,367	18,190
55 to 64	6,489	11,898	15,640	24,410	18,931	6,393	11,548	20,687	24,862	19,527
65+[a]	6,527	13,056	19,240	22,361	16,871	9,784	15,910	5,121	18,868	15,468
					Difference (Percent)					
22 to 27[a]	34	9	13	-24	-9	72	50	-10	-21	-8
28 to 34	23	-29	-23	-1	-2	55	-2	-31	-8	4
35 to 44	-22	-19	-18	-2	-1	59	-11	-26	-24	-11
45 to 54	-2	-22	-15	-2	-1	-28	-23	-21	0	6
55 to 64	107	-17	-22	-1	-2	23	-45	-40	1	4
65+[a]	121	7	-27	-27	-14	59	-15	459	-29	11

Source: Congressional Budget Office based on data from the March 2005 Current Population Survey and from CBO's long-term microsimulation model.

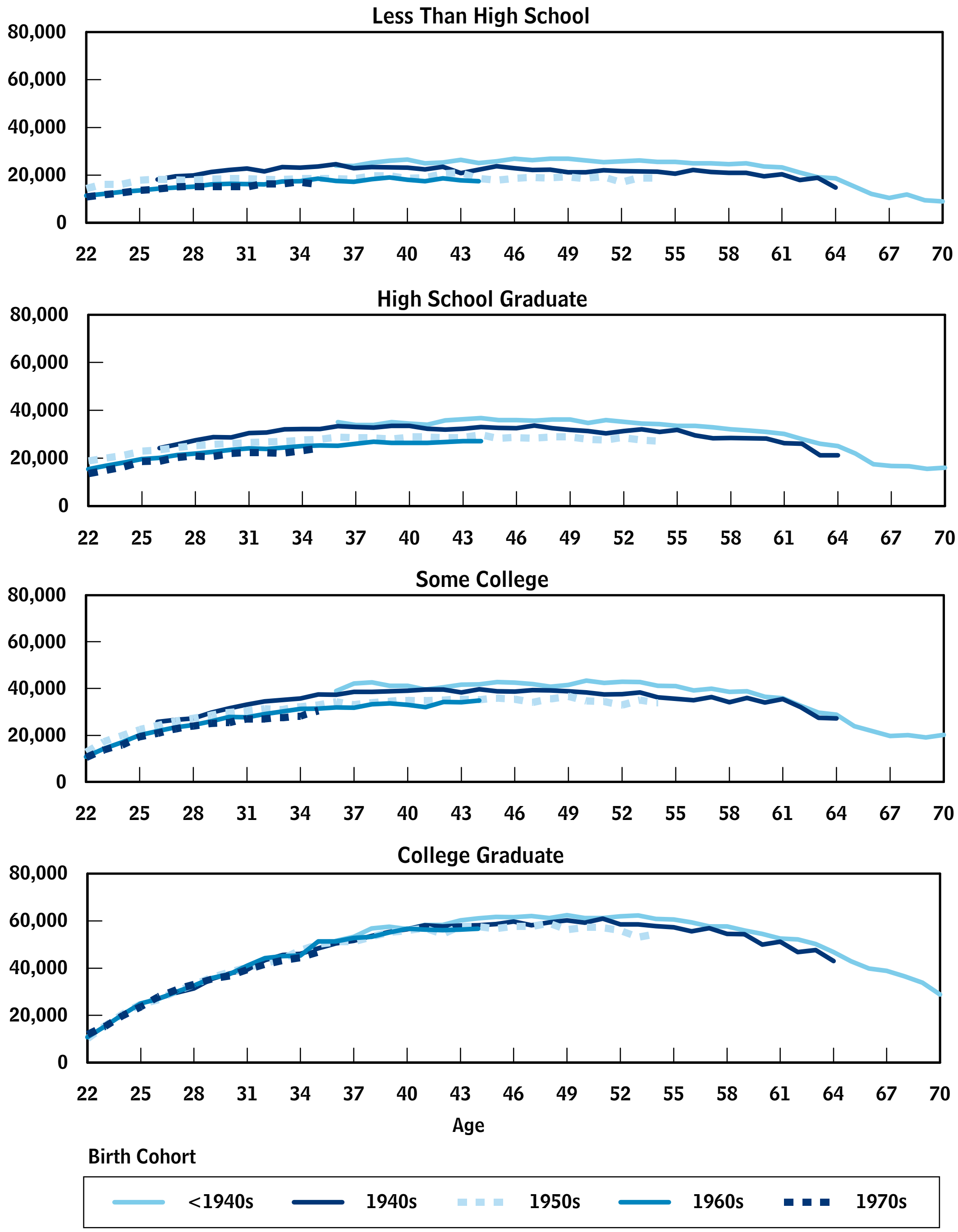
Note: Samples restricted to individuals with earnings above $872 and below $280,000 in 2004. Imputed educational attainment is derived from incorporating estimated regression results from pooled March Current Population Survey files into CBO's long-term microsimulation model. See the text for details.

a. Education status is assigned randomly using pooled Current Population Survey data from 1994 to 2005.

Figure 1.

Earnings for Men, by Educational Attainment, Age, and Birth Cohort

(1993 dollars)

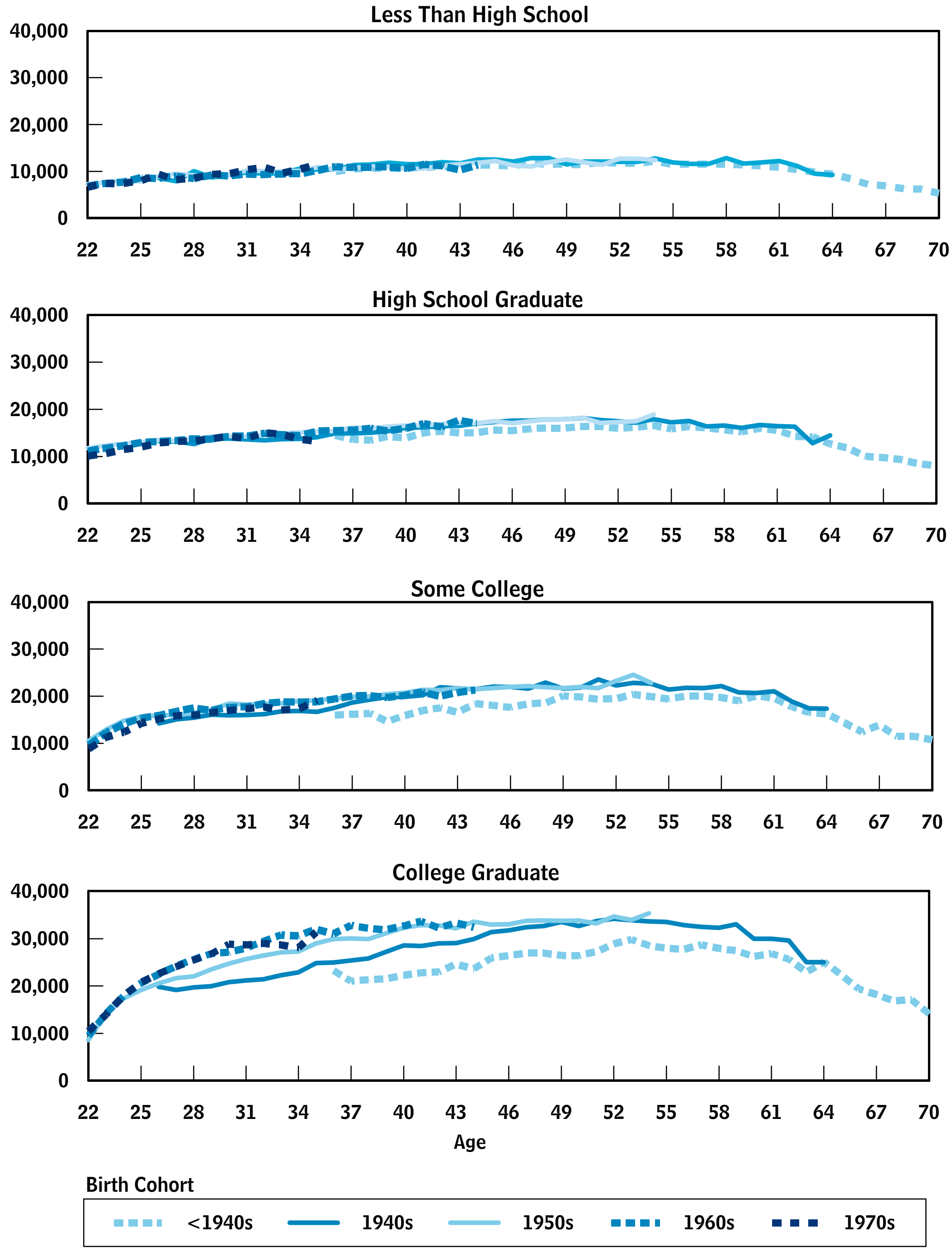


Source: Congressional Budget Office based on data from the March Current Population Survey, 1976 to 2006.

Figure 2.

Earnings for Women, by Educational Attainment, Age, and Birth Cohort

(1993 dollars)

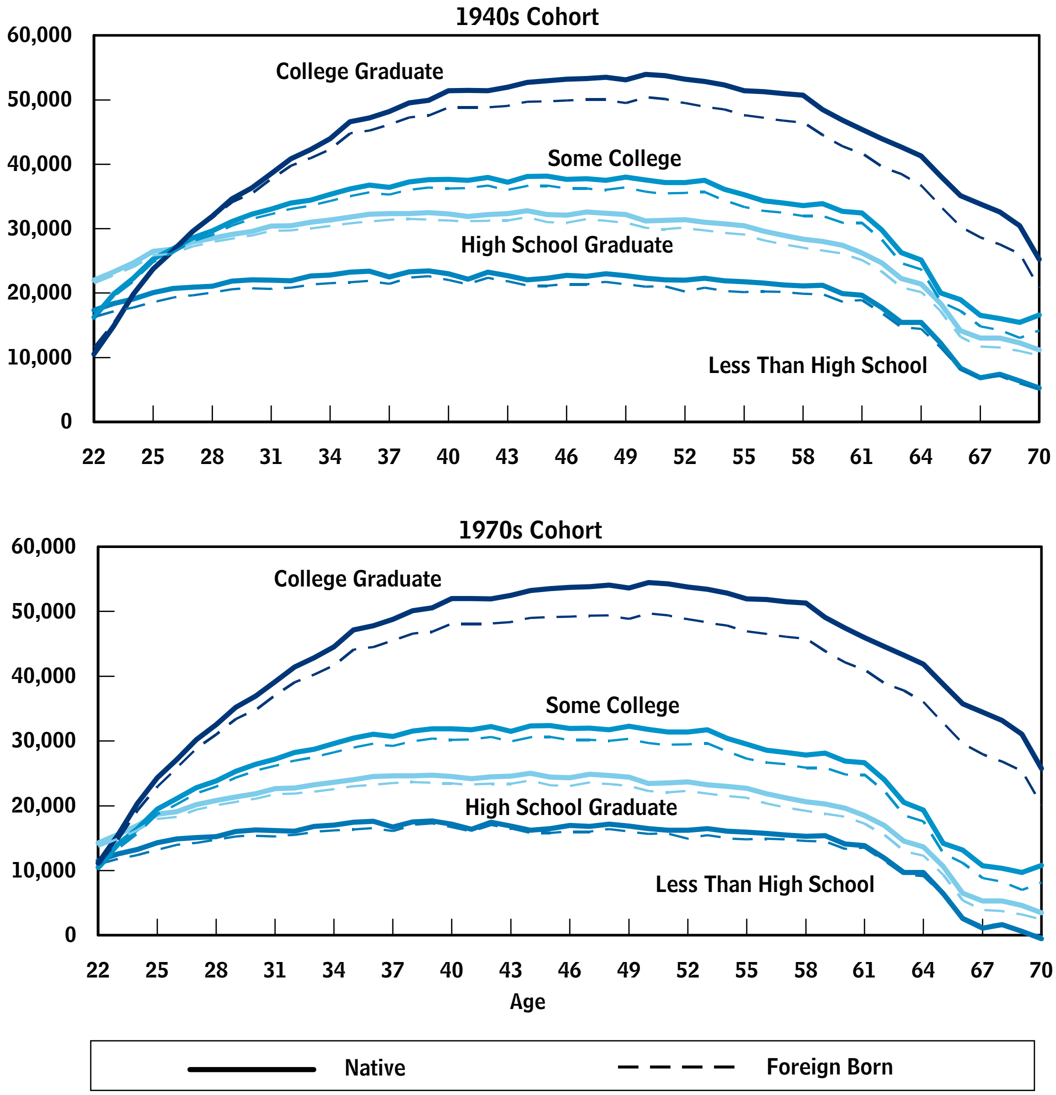


Source: Congressional Budget Office based on data from the March Current Population Survey, 1976 to 2006.

Figure 3.

Predicted Earnings for Men, by Birth Cohort, Age, Educational Attainment, and Nativity Group

(1993 dollars)



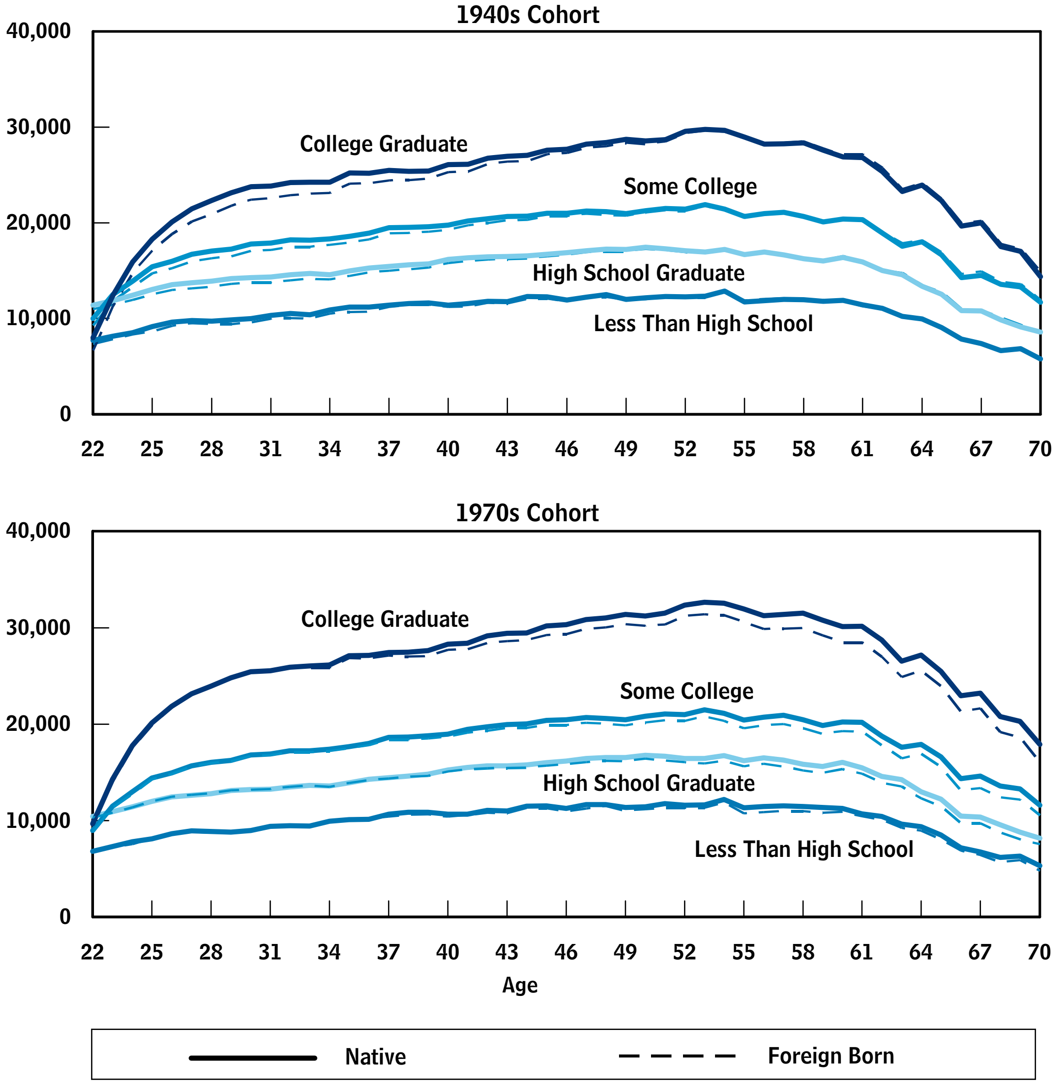
Source: Congressional Budget Office based on data from the March Current Population Survey, 1976 to 2006.

Note: Results were constructed from regressions in which total earnings were predicted using a set of age, education, and cohort dummy variables and a full set of dummy-variable interactions. The regressions were estimated separately for men and women, using different sample weights for native- and foreign-born people. See the text and Appendixes A and B for details.

Figure 4.

Predicted Earnings for Women, by Birth Cohort, Age, Educational Attainment, and Nativity Group

(1993 dollars)



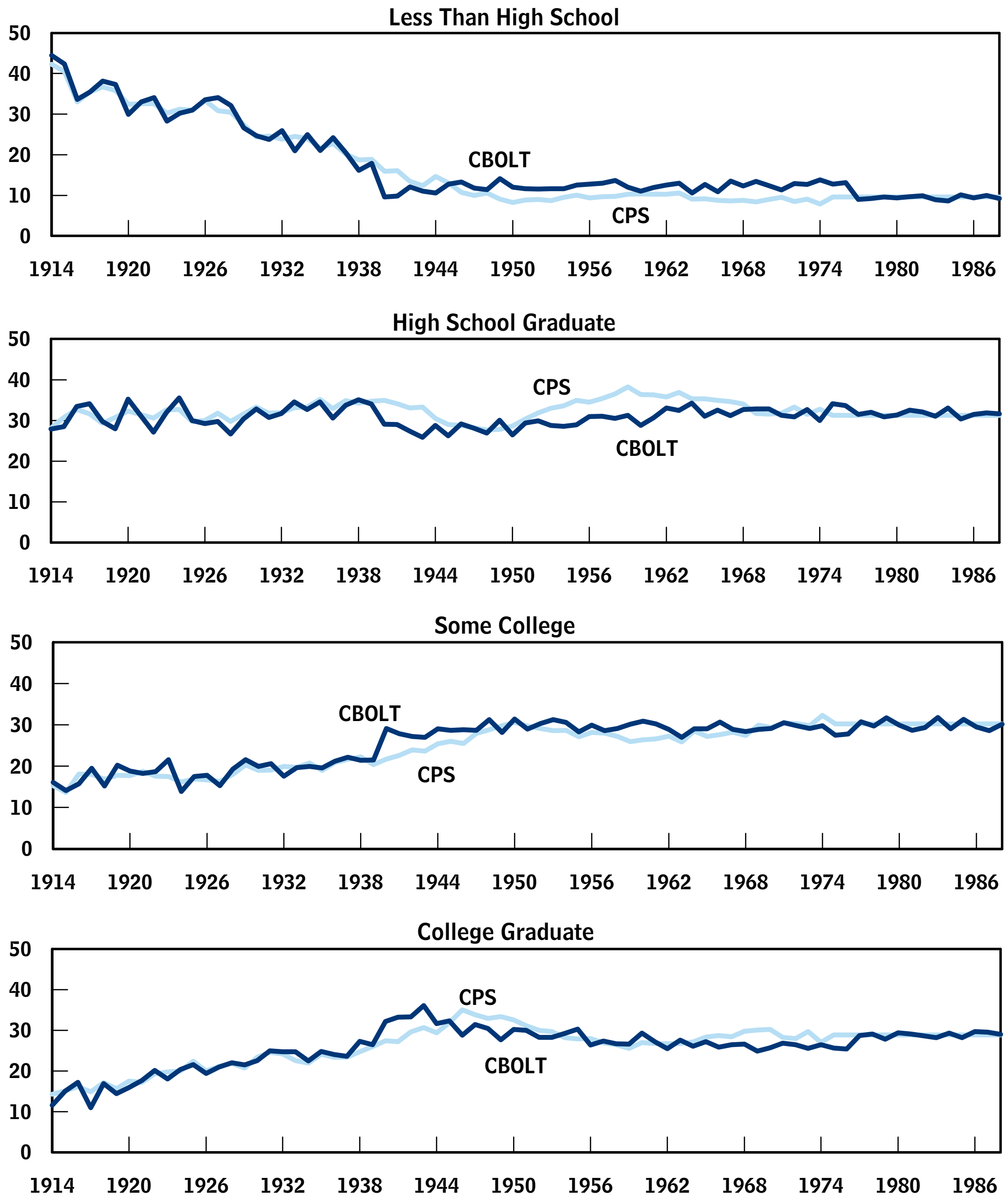
Source: Congressional Budget Office based on data from the March Current Population Survey, 1976 to 2006.

Note: Results were constructed from regressions in which total earnings were predicted using a set of age, education, and cohort dummy variables and a full set of dummy-variable interactions. The regressions were estimated separately for men and women, using different sample weights for native- and foreign-born people. See the text and Appendixes A and B for details.

Figure 5.

Educational Imputation for Native-Born Men, by Birth Year

(Percent of the population)



Source: Congressional Budget Office based on data from the March CPS, 1994 to 2005, and from CBO's long-term microsimulation model.

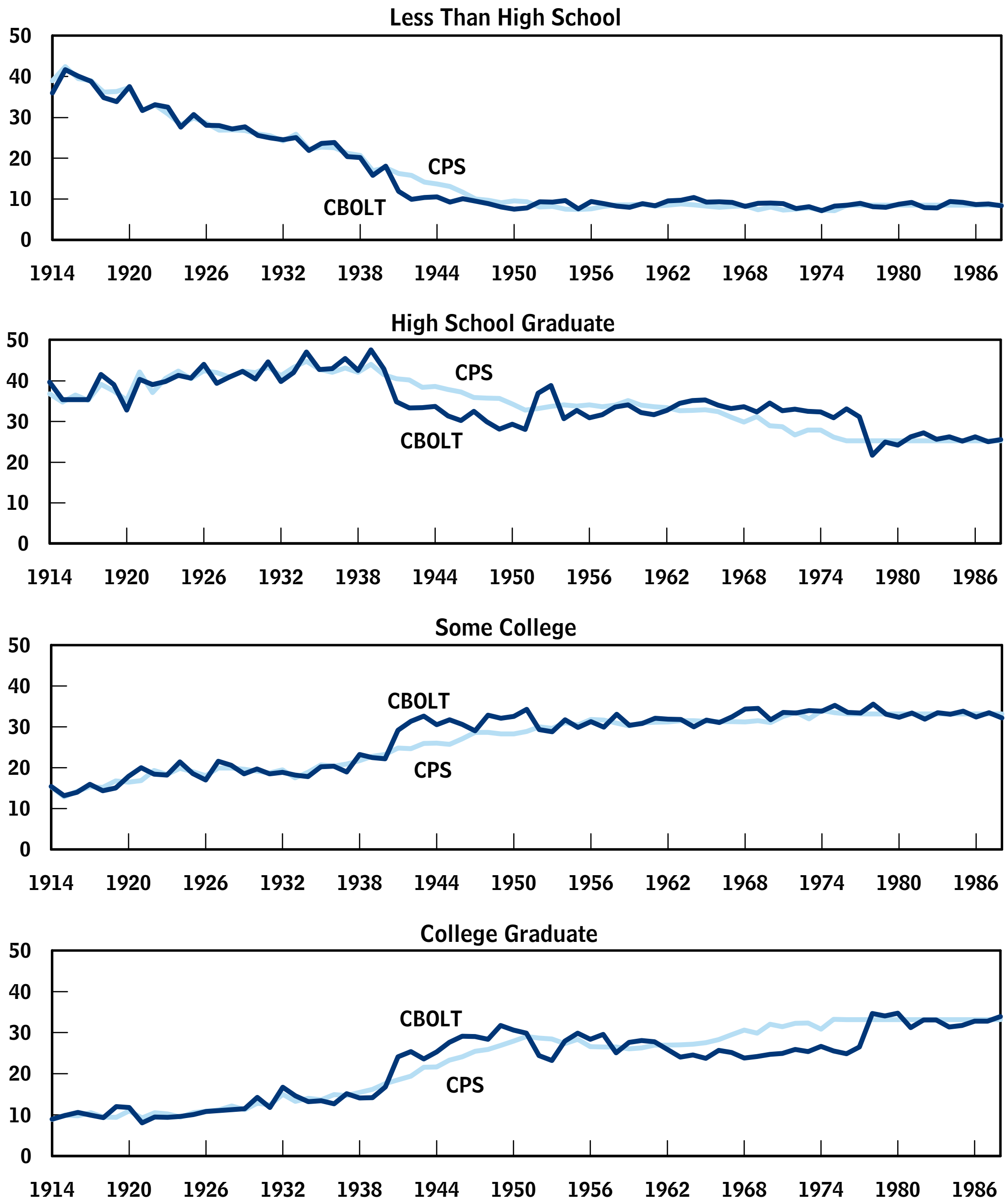
Notes: CPS = Current Population Survey; CBOLT = Congressional Budget Office's long-term microsimulation model.

Imputed educational attainment is derived by incorporating estimated regression results from pooled March CPS files into the CBOLT microsimulation model. See the text for details.

Figure 6.

Educational Imputation for Native-Born Women, by Birth Year

(Percent of the population)



Source: Congressional Budget Office based on data from the March CPS, 1994 to 2005, and from CBO's long-term microsimulation model.

Notes: CPS = Current Population Survey; CBOLT = Congressional Budget Office's long-term microsimulation model.

Imputed educational attainment is derived by incorporating estimated regression results from pooled March CPS files into the CBOLT microsimulation model. See the text for details.

Appendix A: Description of the CPS Data File

The March Current Population Survey (CPS) is a monthly survey of about 50,000 households collected annually by the Census Bureau. The data set contains a wide variety of economic and demographic information on individuals, families, and households in the population. The complete data set the Congressional Budget Office used to create the inputs and estimate the equations in its long-term microsimulation (CBOLT) model combines data on people ages 16 to 90 from the March surveys from 1976 to 2006 (calendar years 1975 to 2005). The full data set consists of more than 2 million observations for the 30-year period.

There are any number of measurement concerns associated with survey data, including, for example, response rates, reporting of multiple jobs held by individual people, and so-called top-coded or imputed earnings (Abowd and Stinson 2007, Cristia and Schwabish 2007, Gottschalk and Huynh 2005, Schwabish 2008).[1] There also are conceptual concerns associated with survey data, including respondents' reports of overtime work, number of hours worked, or other sources of income. Despite those issues, the CPS is widely used and is regarded as a reliable source of labor force data.

Changes in the survey—in collection methods, category designations, and the structure of various components—over the past 30 years could create uneven patterns over time. Before 1994, for example, the data were collected on paper; now, the Census Bureau uses computerized survey instruments (Census Bureau, 2006). Over the years, the conductors of the survey have refined the categories for educational attainment and fine-tuned the way individual earnings are recorded. Those changes could result in more accurate measures of earnings, but proxy response—which occurs when the survey participant reports information about other people—also might introduce bias in the opposite direction (Census Bureau 2006, Roemer 2000, Schwabish 2008).

Educational attainment is now recorded differently than it was 30 years ago. (See Table A-1 for specific categories and the mapping to the definitions used in pooled samples.) In the first half of the CPS sample (before 1991), respondents listed the number of years they spent in formal education. Starting in 1992, respondents were asked about specific milestones of educational attainment, such as "high school graduate–high school diploma" or "12th grade or no diploma."

1. Top-coding of survey data occurs when earnings that exceed some high threshold are capped at a specified limit. Such top-codes often are applied to protect the anonymity of survey participants.

Another difference involved a technical change in the way each component of total earnings was collected and calculated. Before 1988, total earnings were calculated simply as the sum of wage and salary earnings, self-employment earnings, and farm earnings, each top-coded and imputed separately. After 1988, each component must be constructed separately by adding earnings from the respondents' main and other jobs (Burkhauser, Feng, and Jenkins 2007). Again, each component is top-coded and imputed separately.

To account for those and other variations in the survey data, the pooled CPS data were adjusted to create a consistent sample that could be used to estimate the age–earnings profile equations. First, earnings imputed by the Census Bureau for missing survey responses were dropped from the analysis, as suggested by Bollinger and Hirsch (2006) (see also Dooley and Gottschalk 1984). Approximately 5 percent of total earnings are imputed in any given year, although that fraction has grown to more than 10 percent in more recent years.

In the second adjustment, top-coded earnings are multiplied by 1.4, following the process detailed in previous research by, for example, Katz and Murphy (1992) and Lemieux (2007).[2] The level of the top-code, which differs for each of the three earnings components, changed during the sample period, and since 1995, respondents with earnings that exceed a specified top-code had their earnings replaced by age-sex-work experience cell means (see Table A-2). Top-coded earnings in the more recent files are replaced by multiplying the top-code cutoff (for example, $200,000) and not the average value assigned by the Census Bureau.

In the final adjustment, an average wage adjustment indexes total earnings to a common base year. The goal of the earnings equations—to isolate stable earnings patterns across groups—requires that the effects of inflation and productivity in the data be eliminated before estimation. Those effects are then added back in during a simulation, varying appropriately with the state of the macroeconomy. Rather than simply adjusting earnings by means of a standard price index, such as the consumer price index for urban wage earners and clerical workers (CPI-W), earnings are adjusted by using overall average earnings growth, computed as Bureau of Economic Analysis total wages divided by Bureau of Labor Statistics total number of workers. The resulting wage index is highly correlated with the CPI-W because much of the growth in wages over time is attributable to inflation. However, merely adjusting for changes in the CPI-W suggests that $25,000 of earnings in 1993 was worth $32,500 in 2004; the wage index constructed for the CBOLT model suggests that $25,000 in 1993 was worth $37,000 in 2004. That difference in projected value is consistent with the concept that productivity is also important for comparing earnings over long periods.

2. An alternative methodology, as presented in some detail by Lemieux (2007), uses a Pareto distribution to randomly assign earnings above the top-code. However, that methodology created extreme variance in average earnings in some cells, such as older foreign-born workers, for whom there were few observations.

Table A-1.

Education Categories in the Current Population Survey

	1975 to 1990		1991 to 2005	
Category	**Description**	**Code**	**Description**	**Code**
Less Than High School	Elementary	0–11	Less than 1st grade	31
			1st, 2nd, 3rd, or 4th grade	32
			5th or 6th grade	33
			7th or 8th grade	34
	High School	9–12	9th grade	35
			10th grade	36
			11th grade	37
			12th grade or no diploma	38
High School Graduate	High School Diploma	12	High school graduate, high school diploma or equivalent	39
Some College	Some College	13–15	Some college, but no degree	40
			Associate's degree in college, occupation	41
			Associate's degree in college, academic	42
College Graduate	College Degree	16–18	Bachelor's degree	43
			Master's degree	44
			Professional school degree	45
			Doctoral degree	46
Not in Universe	Missing	0	Children, Missing	0

Source: Congressional Budget Office based on data from the March Current Population Survey, 1976 to 2006.

Table A-2.

Top-Coding in the Current Population Survey

(Dollars)

Calendar Years	Income from Wages and Salary	Income from Nonfarm Self-Employment	Income from Farm or Nonincorporated Self-Employment
1975–1980	≥ 50,000	≥ 50,000	≥ 50,000
1981–1986	≥ 75,000	≥ 75,000	≥ 75,000
1987–1994	≥ 99,999	≥ 99,999	≥ 99,999

Calendar Years	Income from Other Wages and Salary	Income from Other Work—Own Business Self-Employment	Income from Other Work—Farm Self-Employment	Earnings from Longest Job Before Deductions
1995–2000	≥ 25,000	≥ 25,000	≥ 40,000	≥ 150,000
2001–2005	≥ 35,000	≥ 35,000	≥ 50,000	≥ 200,000

Source: Congressional Budget Office based on the March Current Population Survey, 1976 to 2006.

Note: Between 1975 and 1994, earnings above the top-code limits are set equal to the top-code. Beginning in 1995, earnings that exceed the top-code limits are set equal to cell averages based on age, sex, and work experience.

Appendix B: Incorporating Immigration Status in the CPS Data File

Beginning with the March 1994 (calendar year 1993) Current Population Survey (CPS), the Census Bureau began to collect information on respondents' place of birth (reported as the nativity group, native or foreign born, to which respondents belong). Because earnings patterns among the foreign born differ from those of native workers, it could be important to estimate age–earnings profiles separately based on place of birth. Unfortunately, restricting the sample to responses from 1993 to 2004 generates too few observations to create smooth age–earnings profiles by age, birth cohort, educational attainment, sex, and nativity group.

A separate imputation procedure to compensate for the lack of nativity data before 1993 uses data from the 1970, 1980, 1990, and 2000 decennial census data files and from the 2004 American Community Survey (ACS), provided by the Integrated Public Use Microdata Series (IPUMS) at the University of Minnesota Population Research Center (Ruggles and others 2004). The IPUMS assigns uniform codes across a variety of different data sets, including the decennial censuses and the ACS. The IPUMS files used in this analysis include the 1970 Form 1 state sample; the 1 percent samples of the 1980, 1990, and 2000 censuses; and the 2004 ACS. All together, the files contain more than 10 million observations.

The adjustments made to the five IPUMS files are similar to those applied to the CPS data: Top-coded earnings are multiplied by 1.4, and nativity is defined by birthplace. To assign nativity in the CPS before 1994, coefficients from a logit regression of foreign born status on age dummy variables, education dummy variables, and real earnings are applied from the IPUMS to the CPS. The regressions are estimated separately for men and women for each year and are weighted using the person-sample weights available in the survey. The pseudo-R^2 statistics from the regressions are between 2 percent and 7 percent, with nearly all coefficients statistically significant at standard levels.

The linear combination of the resulting immigration probabilities between 1970 and 1980, 1980 and 1990, 1990 and 2000, and 2000 and 2004 are then applied to the CPS data to create person-weights that approximate the native and foreign-born populations. Thus, the age–earnings regressions in the CPS are estimated for the same number of survey respondents, but differ by the constructed sample weight.

References

Abowd, John M., and Martha H. Stinson. 2007. "Estimating Measurement Error in SIPP Annual Job Earnings: A Comparison of Census Survey and SSA Administrative Data." Unpublished manuscript. September.

Beaudry, Paul, and David A. Green. 2000. "Cohort Patterns in Canadian Earnings: Assessing the Role of Skill Premia in Inequality Trends." *Canadian Journal of Economics,* vol. 33, no. 4 (November), pp. 907–936.

Bollinger, Christopher, and Barry T. Hirsch. 2006. "Match Bias in the Earnings Imputations in Current Population Survey: The Case of Imperfect Matching." *Journal of Labor Economics*, vol. 24, no. 3 (July), pp. 483–520.

Borjas, George J. 1987. "Self-Selection and the Earnings of Immigrants." *American Economic Review*, vol. 77, no. 4 (September), pp. 531–553.

———. 1989. "Immigrant and Emigrant Earnings: A Longitudinal Study." *Economic Inquiry,* vol. 27, no. 1 (January), pp. 21–37.

Bosworth, Barry, Gary Burtless, and C. Eugene Steuerle. 1999. *Lifetime Earnings Patterns, the Distribution of Future Social Security Benefits, and the Impact of Pension Reform.* Working Paper 1999-06. Chestnut Hill, Mass.: Center for Retirement Research at Boston College. December. crr.bc.edu/images/stories/Working_Papers/wp_1999-06.pdf?phpMyAdmin=43ac483c4de9t51d9eb41.

Bureau of Labor Statistics. 2008. Databases, Tables & Calculators by Subject: Labor Force Statistics from the Current Population Survey. www.bls.gov/data/#employment.

Burkhauser, Richard V., Shuaizhang Feng, and Stephen P. Jenkins. 2007. *Using the P90/P10 Index to Measure U.S. Inequality Trends with Current Population Survey Data: A View from Inside the Census Bureau Vaults.* ISER Working Paper 2007-14. Colchester, U.K.: University of Essex Institute for Social and Economic Research. May. www.iser.essex.ac.uk/pubs/workpaps/pdf/2007-14.pdf.

CBO (Congressional Budget Office). 2006. *Projecting Labor Force Participation and Earnings in CBO's Long-Term Microsimulation Model.* Background Paper. October.

———. 2008. *Recent Trends in the Variability of Individual Earnings and Household Income*. Paper. June.

Census Bureau. 2006. *Design and Methodology: Current Population Survey.* Technical Paper 66. October. www.census.gov/prod/2006pubs/tp-66.pdf.

Cristia, Julian, and Jonathan A. Schwabish. 2007. *Measurement Error in the SIPP: Evidence from Matched Administrative Records.* Congressional Budget Office Working Paper 2007-03. January.

Dooley, Martin D., and Peter Gottschalk. 1984. "Earnings Inequality Among Males in the United States: Trends and the Effect of Labor Force Growth." *Journal of Political Economy,* vol. 92, no.1 (February), pp. 59–89.

Duleep, Harriet Orcutt, and Daniel J. Dowhan. 2008. "Research on Immigrant Earnings." *Social Security Bulletin,* vol. 68, no. 1 (August), pp. 31–50.

Gottschalk, Peter, and Minh Huynh. 2005. *Validation Study of Earnings Data in the SIPP—Do Older Workers Have Larger Measurement Error?* Working Paper 2005-07. Chestnut Hill, Mass.: Center for Retirement Research at Boston College. May. escholarship.bc.edu/cgi/viewcontent.cgi?article =1022&context=retirement_papers.

Juhn, Chinhui, Kevin M. Murphy, and Brooks Pierce. 1993. "Wage Inequality and the Rise in Returns to Skill." *Journal of Political Economy,* vol. 101, no. 3 (June), pp. 410–442.

Katz, Lawrence F., and Kevin M. Murphy. 1992. "Changes in Relative Wages, 1963–1987: Supply and Demand Factors." *Quarterly Journal of Economics,* vol. 107, no. 1 (February), pp. 35–78.

Lemieux, Thomas. 2007. "What Do We Really Know About Changes in Wage Inequality?" Presentation. National Bureau of Economic Research Conference on Research in Income and Wealth, Labor in the New Economy. Bethesda, Md. November 16–17. www.nber.org/confer/2007/CRIWf07/lemieux.pdf.

Lubotsky, Darren. 2007. "Chutes or Ladders? A Longitudinal Analysis of Immigrant Earnings." *Journal of Political Economy,* vol. 115, no. 5 (October), pp. 820–867.

Murphy, Kevin M., and Finis Welch. 1990. "Empirical Age–Earnings Profiles." *Journal of Labor Economics,* vol. 8, no. 2 (April), pp. 202–229.

Panis, Constantijn, and others. 2000. *SSA Program Data User's Manual.* Contract PM-973-SSA. Prepared by RAND Corporation, Santa Monica, Calif., for the Social Security Administration. June.

Passel, Jeffrey S., and Wendy Zimmerman. 2001. *Are Immigrants Leaving California? Settlement Patterns of Immigrants in the Late 1990s.* Washington, D.C.: Urban Institute. April. www.urban.org/url.cfm?ID=410287.

Roemer, Marc. 2000. "Assessing the Quality of the March Current Population Survey and the Survey of Income and Program Participation Income Estimates, 1990–1996." Income Surveys Branch, Housing and Household Economic Statistics Division, Census Bureau. June 16. www.census.gov/hhes/www/income/assess1.pdf.

Ruggles, Stephen, and others. 2004. Integrated Public Use Microdata Series: Version 3.0. Machine-readable database. Minneapolis, Minn.: Minnesota Population Center. usa.ipums.org/usa.

Schwabish, Jonathan A. 2006. *Earnings Inequality and High Earners: Changes During and After the Stock Market Boom of the 1990s.* Congressional Budget Office Working Paper 2006-06. April.

Schwabish, Jonathan A. 2008. "Take a Penny, Leave a Penny: The Propensity to Round Earnings in Survey Data." *Journal of Economic and Social Measurement,* vol. 32, no. 2–3 (January), pp. 93–111.

Schwartz, Amy Ellen, and Benjamin Scafidi. 2004. "What's Happened to the Price of College? Quality-Adjusted Net Price Indexes for Four-Year Colleges." *Journal of Human Resources*, vol. 39, no. 3 (Summer), pp. 723–745.

Swanson, Christopher B. 2008. *Cities in Crisis: A Special Analytic Report on High School Graduation.* Bethesda, Md.: Education Research Center. April 1. www.americaspromise.org/uploadedFiles/AmericasPromiseAlliance/Dropout_Crisis/SWANSONCitiesInCrisis040108.pdf.

CONGRESS OF THE UNITED STATES
CONGRESSIONAL BUDGET OFFICE
WASHINGTON, DC 20515

INSIDE MAIL