APPENDIX A

The family income data series

This appendix explains the various adjustments made to the March Current Population Survey data and the methodology used to prepare the data in the tables discussed on the following pages.

The data source used for our analyses of family incomes and poverty is the U.S. Bureau of the Census's March Current Population Survey (CPS) microdata set. Each March, approximately 60,000 households are asked questions about their incomes from a wide variety of sources in the prior year (the income data in the 2004 March CPS refer to 2003). For the national analysis in Chapter 1, we use the data relevant to the year in question.

In order to preserve the confidentiality of respondents, the income variables on the public-use files of the CPS are top-coded, i.e., values above a certain level are suppressed. Since income inequality measures are sensitive to changes in the upper reaches of the income scale, this suppression poses a challenge to analysts interested in both the extent of inequality in a given time period and the change in inequality over time. We use an imputation technique, described below, that is commonly used in such cases to estimate the value of top-coded data. Over the course of the 1990s, Census top-coding procedures underwent significant changes, which also must be dealt with to preserve consistency. These methods are discussed below.

For most of the years of data in our study, a relatively small share of the distribution of any one variable is top-coded. For example, in 1989, 0.67% (i.e., two-thirds of the top 1%) of weighted cases are top-coded on the variable "earnings from longest job," meaning actual reported values are given for over 99% of those with positive earnings. Nevertheless, the disproportionate influence of the small group of top-coded cases means their earnings levels cannot be ignored.

Our approach has been to impute the average value above the top-code for the key components of income using the assumption that the tails of these distributions follow a Pareto distribution. (The Pareto distribution is defined as c/(x^(a+1)), where c and a are positive constants that we estimate using the top 20% of the empirical distribution (more precisely, c is a scale parameter assumed known; a is the key parameter for estimation). We apply this technique to three key variables: income from wage and salary (1968-1987), earnings from longest job (1988-2000), and income from interest (1968-1992). Since the upper tail of empirical income distributions closely follows the general shape of the Pareto, this imputation method is commonly used for dealing with top-coded data (West, undated). The estimate uses the shape of the upper part of the distribution (in our case, the top 20%) to extrapolate to the part that is unobservable due to the top-codes. Intuitively, if the shape of the observable part of the distribution suggests that the tail above the top-code is particularly long, implying a few cases with very high income values, the imputation will return a high mean relative to the case where it appears that the tail above the top-code is rather short.

Polivka (1998), using an uncensored dataset (i.e., without top-codes), shows that the Pareto procedure effectively replicates the mean above the top-code. For example, her analysis of the use of the technique to estimate usual weekly earnings from the earnings files of the CPS yields estimates that are generally within less than 1% of the true mean.

As noted, the Census Bureau has lifted the top-codes over time in order to accommodate the fact that nominal and real wage growth eventually renders the old top-codes too low. For example, the top-coded value for "earnings from longest job" was increased from $50,000 in 1979 to $99,999 in 1989. Given the growth of earnings over this period, we did not judge this change (or any others in the income-component variables) to create inconsistencies in the trend comparisons between these two time periods.

However, changes made in the mid- and latter 1990s data did require consistency adjustments. For these years, the Census Bureau both adjusted the top-codes (some were raised, some were lowered; the new top-codes were determined by using whichever value was higher: the top 3% of all reported amounts for the variable, or the top 0.5% of all persons), and used "plug-in" averages above the top-codes for certain variables. "Plug-ins" are group-specific average values taken above the top-code, with the groups defined on the basis of gender, race, and worker status. We found that the Pareto procedure was not feasible with unearned income, given the empirical distributions of these variables, so for March data (survey year) 1996 forward we use the plug-in values. Our tabulations show that, in tandem with the procedure described next regarding earnings, this approach avoids trend inconsistencies.

The most important variable that we adjust (i.e., the adjustment with the largest impact on family income) is "earnings from longest job." The top-code on this variable was raised sharply in survey year 1994, and this change leads to an upward bias in comparing estimates at or around that year to earlier years. (Note that this bias is attenuated over time as nominal income growth "catches up" to the new top-code, and relatively smaller shares of respondents again fall into that category.) Our procedure for dealing with this was to impose a lower top-code on the earnings data that we grew over time by the rate of inflation, and to calculate Pareto estimates based on these artificial top-codes. We found that this procedure led to a relatively smooth series across the changes in Census Bureau methodology.

For example, we find that, while our imputed series generates lower incomes among, say, the top 5% of families (because we are imposing a lower top-code) in the mid-1990s, by the end of the 1990s our estimates were only slightly lower than those from the unadjusted Census data. For 2001 forward we do not have any top-code adjustments.

Table 1.2. We decompose the growth of average family income in the following manner. We begin with log changes in family income over the relevant time periods—this is the value to be decomposed between annual hours, hourly wages, and other (non-labor) income. For example, in Table 1.2, this equals 12.8% for the 1994-2000 period. Family earnings grew 15.2% over this period, and we multiply this value by earnings/income averaged over the two years. For this period, that ratio is 0.813. This result represents the earnings contribution (12.4%). In order to decompose this value further into the wage and hours shares, we use weights derived from their growth over the period as shown in the table. The wage share, 1994-2000, is thus computed as (9.9%/ 15.2%)*12.4%, or 8.1%. The share of income growth attributed to the change in "other" is derived by multiplying its growth over the period by the ratio of other/income, again averaged over the two years (note that this is simply one minus the 0.813 value noted above). It is the nature of this type of log decomposition that if the "other" category is a relatively large share of the total, the decomposition will not perfectly sum to the total, but this is not the case here.

Tables 1.26–1.28: The source for these tables is the March CPS datasets described above. The analysis focuses on married-couple families with children, spouse present, where both spouses were between 25 and 54 years of age. The distributional analysis places 20% of families, not persons, in each fifth.

The annual hours variable in the March data is the product of two variables: weeks worked per year, and usual hours per week. Since allowable values on the latter variable go up to 99, this product can be over 5,000. Such values are clearly outliers, and we decided to exclude cases with annual hours greater than 3,500, which led to the exclusion of between 2% and 5% of cases over the years of our analysis.

Wives' wages in this analysis (Table 1.28) are constructed differently than in most of the analysis in this book, i.e., they are "hour-weighted" in this section and "person-weighted" elsewhere. Whereas we usually calculate averages by summing the wages and dividing by the weighted number of earners, in this case we calculate annual hours by dividing annual earnings by annual hours. Since earnings levels and number of hours worked are positively correlated, hour-weighted wage levels tend to be slightly higher than person-weighted wages.

Table 5.10: The methodology for this decomposition is taken from Danziger and Gottschalk (1995, chapter 5). The change to be explained is the difference in poverty rates between t_0 and t_1. We first isolate the effect of average income growth by assigning the average growth between the two time periods to all families in t_0 and recalculate the poverty rate (we adjust each family's poverty line for the increase in the CPI over this period). This procedure holds the demographic composition and the shape of the income distribution constant in t_0 while allowing incomes to grow equally for all families. Thus, the difference between this simulated poverty rate and the actual t_0 poverty rate is attributable to the growth in average income.

We repeat this exercise for each demographic group in t_0 (we use the three family types in Table 5.8, two races—white and non-white—and three education categories of the family head—less than high school, high school and some college, and college or more). By weighting each of these simulated t_0 rates by their t_1 population shares, we can simulate a t_0 poverty rate that reflects the average income growth and demographic composition of t_1. The difference between this simulated rate and the one discussed in the above paragraph gives the contribution of demographic change over the time period. Finally, since this second simulated rate incorporates the mean growth and demographic change between the two periods, but not the change in the shape of the distribution, the difference between this second simulated rate and the actual rate for t_1 equals the change in poverty rates attributable to changes in inequality over the two periods.