The State of Working America
12th Edition

LAWRENCE MISHEL • JOSH BIVENS
ELISE GOULD • HEIDI SHIERHOLZ

Appendices

EPI DIGITAL EDITION

This chapter is from *The State of Working America, 12th Edition,* an Economic Policy Institute book published by Cornell University Press in November 2012.

Data from this chapter should be attributed to the Economic Policy Institute's *The State of Working America, 12th Edition*.

APPENDIX A

Appendix A
CPS income measurement

This appendix explains the various adjustments made to microdata from the U.S. Census Bureau's Current Population Survey Annual Social and Economic Supplement (CPS-ASEC, commonly referred to as the March Supplement or March CPS) and the methodology used to prepare the data. The CPS is a monthly survey of unemployment and labor force participation prepared by the U.S. Census Bureau for the Bureau of Labor Statistics, and the CPS-ASEC is a special annual questionnaire that gathers income and earnings data. The microdata are raw untabulated survey responses. This microdata set is one of the data sources used for our analyses of family and household incomes as well as poverty. Each March, approximately 60,000 households are asked questions about their incomes from a wide variety of sources in the prior year (for example, the income data in the 2011 March CPS refer to 2010).

In order to preserve the confidentiality of respondents, the income variables in the public-use files of the CPS are top-coded, that is, values above a certain level are capped at a single common value. The reasoning is that since so few individuals, if any, have incomes above this "top-code," reporting the *exact* income number could allow somebody to use that information (along with other information from the March CPS, such as state of residence, age, ethnicity, etc.) to actually identify a specific survey respondent. Since income inequality measures are sensitive to changes in the upper reaches of the income distribution, this top-coding poses a challenge to analysts interested in both the extent of inequality in a given 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 percent (i.e., two-thirds) of the top 1 percent of weighted cases are top-coded on the variable "earnings from longest job," meaning actual reported values are given for more than 99 percent 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 percent 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 distribution, this imputation method is commonly used for dealing with top-coded data. The estimate uses the shape of the upper part of the distribution (in our case, the top 20 percent) 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 in which the tail above the top-code appears rather short.

Polivka and Miller (1998), using an uncensored dataset (i.e., without top-codes), show that the Pareto procedure effectively replicates the mean above the top-code. For example, Polivka and Miller's 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 percent of the true mean.

As noted, the U.S. 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 years.

However, changes made in the mid- to late 1990s data did require consistency adjustments. For these years, the Census Bureau adjusted the top-codes: Some were raised, some were lowered, and the new top-codes were determined by using the higher value of either the top 3 percent of all reported amounts for the variable or the top 0.5 percent of all persons. The bureau also 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 from (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, to grow that top-code 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 percent 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.

Appendix B
Wage measurement

This appendix provides background information on the analysis of wage data from the Current Population Survey (CPS), which is best known for providing the monthly estimates of unemployment. The CPS is prepared by the U.S. Census Bureau for the Bureau of Labor Statistics (BLS). Specifically, for 1979 and beyond, we analyze microdata files that contain a full year's data on the outgoing rotation groups (ORG) in the CPS. (For years prior to 1979, we use the CPS May files; our use of these files is discussed later in this appendix.) We believe that the CPS-ORG files allow for timely and accurate analyses of wage trends that are in keeping with the familiar labor force definitions and concepts employed by BLS.

The sampling framework of the monthly CPS is a "rolling panel," in which households are in the survey for four consecutive months, out for eight, and then back in for four months. The ORG files provide data on those CPS respondents in either the fourth or eighth month of the CPS (i.e., in groups four or eight, out of a total of eight groups). Therefore, in any given month, the ORG file represents a quarter of the CPS sample. For a given year, the ORG file is equivalent to three months of CPS files (one-fourth of 12 months). For our analyses, we use a sample drawn from the full-year ORG sample, the size of which ranges from 160,000 to 180,000 observations during the 1979 to 1995 period. Due to a decrease in the overall sample size of the CPS, the ORG shrank to 145,000 cases from 1996 to 1998, and our most recent sample contains about 170,000 cases.

Changes in annual or weekly earnings can result from changes in hourly earnings or changes in time worked (hours worked per week or weeks worked per year). Our analyses focus on the hourly wage, which represents the pure price of labor (exclusive of benefits), because we are interested in changing pay levels for the workforce and its subgroups. This enables us to clearly distinguish changes in earnings resulting from more (or less) work from changes resulting from more (or less) pay. Most of our wage analyses, therefore, do not account for weekly or annual earnings changes due to reduced or increased work hours or opportunities for employment. An exception is Table 4.1, which presents annual hours, earnings, and hourly wages from the March CPS and shows that the overwhelming driver of annual wage trends between business cycle peaks has been trends in hourly wages.

In our view, the ORG files provide a better source of data for wage analyses than the traditionally used March CPS files. In order to calculate hourly wages from the March CPS, analysts must make calculations using three retrospective variables: the annual earnings, weeks worked, and usual weekly hours worked in the year prior to the survey. In contrast, respondents in the ORG are asked a set of questions about hours worked, weekly wages, and, for workers paid by the hour, hourly wages in the week prior to the survey. In this regard, the data from the ORG are likely to be more reliable than data from the March CPS. See Bernstein and Mishel (1997) for a detailed discussion of these differences.

Our subsample includes all wage and salary workers with valid wage and hour data, whether paid weekly or by the hour. Specifically, in order to be included in our subsample, respondents had to meet the following criteria:

- age 18–64
- employed in the public or private sector (unincorporated self-employed were excluded)
- hours worked within the valid range in the survey (1–99 per week, or hours vary—see discussion below)
- either hourly or weekly wages within the valid survey range (top-coding discussed below)

For those who met these criteria, an hourly wage was calculated in the following manner: If a valid hourly wage was reported, that wage was used throughout our analysis. For salaried workers (those who report only a weekly wage), the hourly wage was their weekly wage divided by their hours worked. Outliers, i.e., persons with hourly wages below 50 cents or above $100 in 1989 dollars (adjusted by the CPI-U-X1 consumer price index), were removed from the analysis. Starting from year 2002, we use dollars adjusted by the Consumer Price Index Research Series Using Current Methods (CPI-U-RS). These yearly upper and lower bounds are presented in **Table B.1**. CPS demographic weights were applied to make the sample nationally representative.

The hourly wage reported by hourly workers in the CPS excludes overtime, tips, or commissions (OTTC), thus introducing a potential undercount in the

Table B.1 Wage earner sample, hourly wage lower and upper limits, 1973–2011

	Lower	Upper
1973	$0.19	$38.06
1974	0.21	41.85
1975	0.23	45.32
1976	0.24	47.90
1977	0.25	50.97
1978	0.27	54.44
1979	0.30	59.68
1980	0.33	66.37
1981	0.36	72.66
1982	0.39	77.10
1983	0.40	80.32
1984	0.42	83.79
1985	0.43	86.77
1986	0.44	88.39
1987	0.46	91.61
1988	0.48	95.40
1989	0.50	100.00
1990	0.53	105.40
1991	0.55	109.84
1992	0.57	113.15
1993	$0.58	$116.53
1994	0.60	119.52
1995	0.61	122.90
1996	0.63	126.53
1997	0.65	129.54
1998	0.66	131.45
1999	0.67	134.35
2000	0.69	138.87
2001	0.71	142.82
2002*	0.70	140.05
2003*	0.72	143.26
2004*	0.74	147.06
2005*	0.76	152.10
2006*	0.78	156.90
2007*	0.81	161.45
2008*	0.84	167.66
2009*	0.84	167.04
2010*	0.85	169.78
2011*	0.88	175.14

* Upper limit adjusted by CPI-U-RS

Source: Authors' analysis of Current Population Survey Outgoing Rotation Group microdata

hourly wage for workers who regularly receive tips or premium pay. OTTC is included in the usual weekly earnings of hourly workers, which raises the possibility of assigning an imputed hourly wage to hourly workers based on the reported weekly wage and hours worked per week. Conceptually, using this imputed wage is preferable to using the reported hourly wage because it is more inclusive. We have chosen, however, not to use this broader wage measure, because the extra information on OTTC seems unreliable. We compared the imputed hourly wage (reported weekly earnings divided by weekly hours) to the reported hourly wage; the difference presumably reflects OTTC. This comparison showed that significant percentages of the hourly workforce appeared to receive negative OTTC. These error rates range from a low of 0 percent of the hourly workforce in 1989–1993 to a high of 16–17 percent in 1973–1988, and persist across the survey change from 1993 to 1994. Since negative OTTC is clearly implausible, we rejected this imputed hourly wage series and rely strictly on the hourly rate of pay as reported directly by hourly workers, subject to the sample criteria discussed above.

For tables that show wage percentiles, we "smooth" hourly wages to compensate for "wage clumps" in the wage distributions. The technique involves creating a categorical hourly wage distribution, where the categories are 50-cent intervals, starting at 25 cents. We then find the categories on either side of each decile and perform a weighted, linear interpolation to locate the wage precisely on the particular decile. The weights for the interpolation are derived from differences in the cumulative percentages on either side of the decile. For example, suppose that 48 percent of the wage distribution of workers by wage level are in the $9.26–$9.75 wage "bin," and 51 percent are in the next higher bin, $9.76–$10.25. The weight for the interpolation (in this case, the median, or 50th percentile) is (50–48)/(51–48), or two-thirds. The interpolated median equals this weight, times the width of the bin ($.50), plus the upper bound of the previous bin ($9.75); $10.08 in this example.

In order to preserve the confidentiality of respondents, the income variables in the public-use files of the CPS are top-coded, that is, values above a certain level are capped at a single common value. The reasoning is that since so few individuals, if any, have incomes above this "top-code," reporting the exact income number could allow somebody to use that information (along with other information from the CPS, such as state of residence, age, ethnicity, etc.) to actually identify a specific survey respondent. For the survey years 1973–1985, the weekly wage is top-coded at $999.00; an extended top-code value of $1,923 is available in 1986–1997; the top-code value changes to $2,884.61 in 1998 and remains at that level. Particularly for the later years, this truncation of the wage distribution creates a downward bias in the mean wage. We dealt with the top-coding issue by imputing a new weekly wage for top-coded individuals. The imputed value is the Pareto-imputed mean for the upper tail of the weekly earnings distribution, based on the distribution of weekly earnings up to the 80th percentile (see Appendix A for a discussion of the Pareto distribution). This procedure was done for men and women separately. The imputed values for men and women appear in **Table B.2**. A new hourly wage, equal to the new estimated value for weekly earnings, divided by that person's usual hours per week, was calculated.

In January 1994, a new survey instrument was introduced into the CPS; many labor force items were added and improved. This presents a significant challenge to researchers who wish to make comparisons over time. The most careful research on the impact of the survey change has been conducted by BLS researcher Anne Polivka (1996). Interestingly, Polivka did not find that the survey changes had a major impact on broad measures of unemployment or wage

Table B.2 Pareto-imputed mean values for top-coded weekly earnings, and share top coded, 1973–2011 *Part 1 of 2*

	Share (percent hours)			**Value**	
	All	**Men**	**Women**	**Men**	**Women**
1973	0.11%	0.17%	0.02%	$1,365	$1,340
1974	0.16	0.26	0.01	1,385	1,297
1975	0.21	0.35	0.02	1,410	1,323
1976	0.30	0.51	0.01	1,392	1,314
1977	0.36	0.59	0.04	1,384	1,309
1978	0.38	0.65	0.02	1,377	1,297
1979	0.57	0.98	0.05	1,388	1,301
1980	0.72	1.23	0.07	1,380	1,287
1981	1.05	1.82	0.10	1,408	1,281
1982	1.45	2.50	0.18	1,430	1,306
1983	1.89	3.27	0.25	1,458	1,307
1984	2.32	3.92	0.42	1,471	1,336
1985	2.78	4.63	0.60	1,490	1,343
1986	0.80	1.37	0.15	2,435	2,466
1987	1.06	1.80	0.20	2,413	2,472
1988	1.30	2.19	0.29	2,410	2,461
1989	0.48	0.84	0.08	2,710	2,506
1990	0.60	1.04	0.11	2,724	2,522
1991	0.71	1.21	0.17	2,744	2,553
1992	0.77	1.28	0.22	2,727	2,581
1993	0.86	1.43	0.24	2,754	2,580
1994	1.25	1.98	0.43	2,882	2,689
1995	1.34	2.16	0.43	2,851	2,660
1996	1.41	2.27	0.46	2,863	2,678
1997	1.71	2.67	0.65	2,908	2,751
1998	0.63	0.98	0.25	4,437	4,155
1999	0.71	1.12	0.21	4,464	4,099
2000	0.83	1.38	0.24	4,502	4,179
2001	0.92	1.46	0.34	4,477	4,227
2002	0.91	1.44	0.33	4,555	4,252
2003	1.07	1.69	0.40	4,546	4,219
2004	1.19	1.90	0.42	4,611	4,195
2005	1.30	2.02	0.51	4,623	4,264
2006	1.49	2.26	0.65	4,636	4,328
2007	1.69	2.55	0.76	4,658	4,325

Table B.2 **Pareto-imputed mean values for top-coded weekly earnings, and share top coded, 1973–2011** ***Part 2 of 2***

	Share (percent hours)			**Value**	
	All	**Men**	**Women**	**Men**	**Women**
2008	1.95%	2.95%	0.87%	$4,723	$4,383
2009	2.09	3.20	0.92	4,872	4,403
2010	2.25	3.33	1.11	4,888	4,458
2011	2.26	3.30	1.14	4,792	4,477

Source: Authors' analysis of Current Population Survey Outgoing Rotation Group microdata

levels, though significant differences did surface for some subgroups (e.g., weekly earnings for those with less than a high school diploma and those with advanced degrees, and the unemployment rate of older workers). However, a change in the reporting of weekly hours did call for the alteration of our methodology. In 1994 the CPS began allowing people to report that their usual hours worked per week vary. In order to include nonhourly workers who report varying hours in our wage analyses, we estimated their usual hours using a regression-based imputation procedure, in which we predicted the usual hours of work for "hours vary" cases based on the usual hours worked of persons with similar characteristics. An hourly wage was calculated by dividing weekly earnings by the estimate of hours for these workers. The share of our sample that received such a wage in the 1994–2005 period is presented in **Table B.3**. The reported hourly wage of hourly workers was preserved.

BLS analysts Ilg and Haugen (2000), following Polivka (2000), did adjust the 10th-percentile wage because "changes to the survey in 1994 led to lower reported earnings for relatively low-paid workers, compared with pre-1994 estimates." We make no such adjustments for both practical and empirical reasons. Practically, the BLS has provided no adjustment factors for hourly wage trends that we can use—Polivka's work is for weekly wages. More important, the trends in 10th-percentile hourly wages differ from those reported by Ilg and Haugen for 10th-percentile weekly earnings. This is perhaps not surprising, since the composition of earners at the "bottom" will differ when measured by weekly rather than hourly wages, with low-weekly earners being almost exclusively part-timers. Empirically, Ilg and Haugen show the unadjusted 50/10 wage gap increasing between 1993 and 1994, when the new survey begins. In contrast, our 50/10 wage gap for hourly wages decreases between 1993 and 1994. Thus, the pattern of wage change in their data differs greatly from that in our data. In fact, our review of the 1993–1994 trends across all of the deciles shows no discontinuities whatsoever. Consequently, we make no adjustments to account for any effect of

Table B.3 Share of wage earners assigned an hourly wage from imputed weekly hours, 1994–2011

	Percent hours vary
1994	2.0%
1995	2.1
1996	2.4
1997	2.4
1998	2.5
1999	2.4
2000	2.4
2001	2.5
2002	2.5
2003	2.5
2004	2.7
2005	2.7
2006	2.5
2007	2.4
2008	2.4
2009	2.3
2010	2.1
2011	2.0

Source: Authors' analysis of Current Population Survey Outgoing Rotation Group microdata

the 1994 survey change. Had we made the sort of adjustments suggested by Polivka, our measured fall in the 50/10 wage gap in the 1990s would be even larger, and the overall pattern—wage gaps shrinking at 50/10, widening at 90/50, and, especially, at 95/50—would remain the same.

When a response is not obtained for weekly earnings, or an inconsistency is detected, an "imputed" response is performed by CPS using a "hot deck" method, whereby a response from another sample person with similar demographic and economic characteristics is used for the nonresponse. This procedure for imputing missing wage data appears to bias comparisons between union and nonunion members. We restrict our sample to the observations with non-imputed wages only for analysis of the union wage premium (Table 4.33).

Racial/ethnic demographic variables are also used in tables and in results reporting wage regression analyses. Starting in January of 2003, individuals are asked directly if they belong to Spanish, Hispanic, or Latino categories. Persons who report they are Hispanic also may select more than one race. For consistency, our race variable includes four mutually exclusive categories across years:

- white, non-Hispanic
- black, non-Hispanic
- Hispanic, any race
- all others

In January 2003, the CPS used the 2002 Census Bureau occupational and industry classification systems, which are derived from the 2000 Standard Occupational Classification (SOC) system and the 2002 North American Industry Classification System (NAICS). The new classification systems create breaks in existing data series at all levels of aggregation. Since we have built in "old" and "new" industry and occupation systems in our underlying 2000–2002 data, we use year 2000 as a break point to create consistent analyses with the "old" code for pre-2000 analysis and the "new" code for post-2000 analysis.

Beginning in 1992, the CPS employed a new coding scheme for education, providing data on respondents' highest degree attained. In earlier years, the CPS provided data on years of schooling completed. The challenge of making a consistent wage series by education level is to either make the new data consistent with the past or to make the old "years of schooling" data consistent with the new educational attainment measures. In prior editions of *The State of Working America*, we achieved a consistent series by imputing years of schooling for 1992 and later years, i.e., making the "new" consistent with the "old." In this version, however, we have converted the old data to the new coding following Jaeger (1997). However, Jaeger does not separately identify four-year college and "more than college" categories. Since the wages of these subgroups of the "college or more" group have divergent trends, we construct pre-1992 wages and employment separately for "four-year college" and "advanced." To do so, we compute wages, wage premiums, and employment separately for those with 16, 17, and 18-plus years of schooling completed. The challenge is to distribute the "17s" to the 16 years (presumably a four-year degree) and 18-plus years (presumably advanced) groups. We do this by using the share of the 17s that have a terminal four-year college degree, as computed in the February 1990 CPS supplement that provides both education codings: 61.4 percent. We then assume that 61.4 percent of all of the 17s are "college only" and compute a weighted average of the 16s and 61.4 percent of the 17s to construct "college only" wages and wage premiums. Correspondingly, we compute a weighted average of 38.6 percent (or 1 less 61.4 percent) of the 17s and the 18s to construct advanced "wages and wage premiums." Distributing the 17s affects each year differently depending on the actual change in the wages and premiums for 17s and the changing relative size of the 17s (which varies only slightly from 2.5 percent of men and women from 1979 to 1991).

We employ these education categories in various tables in Chapter 4, where we present wage trends by education over time. For the data for 1992 and later, we compute the "some college" trends by aggregating those "with some college but no degree beyond high school" and those with an associate or other degree that is not a four-year college degree.