NBER WORKING PAPER SERIES

USING THE PARETO DISTRIBUTION TO IMPROVE ESTIMATES OF TOPCODED EARNINGS

Philip Armour
Richard V. Burkhauser
Jeff Larrimore

Working Paper 19846
http://www.nber.org/papers/w19846

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
January 2014

Armour over the past 12 months has received funding in excess of $5,000 from the Disability Research Consortium, and funding not in excess of $5,000 from the Association for Convenience and Fuel Retailing. Burkhauser over the past 12 months has received funding in excess of $5,000 from the National Institute on Disability and Rehabilitation Research and the Employment Policies Institute. In addition he received funding not in excess of $5,000 from: The American Enterprise Institute, The Brookings Institution, the Federal Reserve Board, the Fraser Institute, the National Institute on Aging, the Pew Charitable Trusts, and the Russell Sage Foundation. Larrimore over the past 12 months has received funding not in excess of $5,000 from the Russell Sage Foundation and the Fraser Institute. Support for this research from the National Science Foundation (award nos. SES-0427889 SES-0322902, and SES-0339191) and the Employment Policy and Measurement Rehabilitation Research and Training Center at the University of New Hampshire, which is funded by the National Institute for Disability and Rehabilitation Research (NIDRR, grant no. H133B100030) are cordially acknowledged. The research in this paper was conducted while the authors were Special Sworn Status researchers of the U.S. Census Bureau at the New York Census Research Data Center at Cornell University. This paper has been screened to ensure that no confidential data are disclosed. The views and opinions expressed herein are those of the authors and should not be attributed to the Joint Committee on Taxation, any Member of Congress, the Census Bureau, or the National Bureau of Economic Research.

Using the Pareto Distribution to Improve Estimates of Topcoded Earnings
Philip Armour, Richard V. Burkhauser, and Jeff Larrimore
NBER Working Paper No. 19846
January 2014
JEL No. C81,D31,J01,J31

ABSTRACT

Inconsistent censoring of top earnings in the public-use March Current Population Survey (CPS) is an important limitation in using it to measure labor earnings trends. Using less-censored internal CPS data, combined with Pareto estimates from it for internally censored observations, we create an enhanced cell-mean series to capture top earnings in the public-use CPS. We find previous common approaches for imputing topcoded earnings systematically understate top earnings. Annual earnings inequality trends since 1963 using our series closely approximate the substantial increase in earnings inequality observed in Social Security Administration data for working-age commerce and industry workers by Kopczuk, Saez, and Song (2010). However, when considering all workers the level of earnings inequality is higher but the increase over this time has been more modest.

Philip Armour
Cornell University
404 Uris Hall
Ithaca, NY 14853
poa8@cornell.edu

Jeff Larrimore
Joint Committee on Taxation
1625 Longworth House Office Building
Washington, D.C. 20515
jeff.larrimore@mail.house.gov

Richard V. Burkhauser
Cornell University
Department of Policy Analysis & Management
259 MVR Hall
Ithaca, NY 14853-4401
and University of Melbourne
and also NBER
rvb1@cornell.edu

The public-use March Current Population Survey (CPS) is the primary source of data for tracking levels and trends in United States labor earnings and labor earnings inequality, and for explaining their causes. This literature has especially focused on whether the rise in wage and earnings inequality in the 1980s was part of a long-run secular trend or an episodic event (Autor, Katz, & Kearney, 2008; Card & DiNardo, 2002; Juhn, Murphy, & Pierce, 1993. See Acemoglu, 2002, for a review of this literature). However, this public-use CPS-based literature has been hampered by its attenuated view of the right tail of the labor earnings distribution due to the topcoding of high earnings in the public-use CPS data.[1]

To correct for topcoding biases, CPS-based researchers have generally pursued one of three paths: (1) ignoring the topcoding problem; (2) making an ad-hoc adjustment to topcoded earnings values; or (3) using a Pareto distribution to estimate earnings at the top of the distribution. For example, a common ad-hoc technique, based on estimates from Pareto imputations of top earnings, is to replace topcoded earnings with a multiple of the topcode threshold, so all individuals with topcoded earnings in a year are assumed to have earnings at 1.3, 1.4, or 1.5 times the topcode threshold (Autor, Katz, & Kearney, 2008; Katz & Murphy, 1992; Juhn, Murphy, & Pierce, 1993; Lemieux, 2006). However, such an approach may misstate top earnings if the wrong multiple is used or if the appropriate multiple changes over time. Similarly, researchers using a Pareto imputation of top earnings may misstate those earnings if they are unable to obtain a reasonable fit for the Pareto distribution when using available public-use data.

Making use of internal March CPS files with their much higher censoring levels, we show that previous ad-hoc estimates and Pareto estimations of top earnings based on public-use

[1] Some wage inequality research focuses on the wage questions in the May Outgoing Rotation Group (ORG) sample of the CPS, which is also subject to topcoding.

data understate mean earnings at the top of the earnings distribution and hence earnings inequality. Then, using a continuous maximum likelihood estimator along with internal CPS data, we produce a series of more accurate estimates of top earnings in the CPS data. Our estimates start with actual top earnings from the internal CPS data combined with a Pareto estimate using these data for internally censored observations. With this hybrid approach, we create an enhanced cell-mean series that allows researchers who only have access to the public-use data to more accurately capture top earnings levels and trends.

To show the value of our new measure, we use it together with the public-use CPS to replicate the level and trend in labor earnings inequality from 1963 to 2004 that Kopczuk, Saez, and Song (2010) find using Social Security (SSA) administrative records for the subsample of U.S. workers who paid social security taxes in the Commerce and Industry sector of the labor market.

II. Data

The March CPS survey contains a comprehensive set of questions on sources of household earnings, including labor earnings.[2] Figure 1 provides an overview of the public topcode and internal censoring levels for annual wage earnings from 1967-1986 and for primary labor earnings, which are primarily wages, from 1987-2007 – which encompasses the range of years for which we have access to the internal CPS data. Both the public topcode level and the internal censoring level (left y-axis) increase on an irregular, ad-hoc basis. As a result the percentage of individuals with earnings above the public topcode (right y-axis) rises steadily when topcodes are held nominally constant and falls quickly when these topcodes are raised.

[2] The March CPS asks about income in the previous year, so the income year is always one year prior to the survey year. All references to years in this paper refer to the income year. Because of Census Bureau changes in their aggregation techniques we use wage and salary earnings for years prior to income year 1987 and all primary labor earning thereafter. Since the vast majority of primary earnings are from wages and salaries, this break does not appear to have a noticeable impact on our results.

III: Estimating Top Earnings

Most researchers who measure long-term trends in earnings with public-use CPS data have adopted ad-hoc techniques to correct for topcoding, such as imputing topcoded earnings as a fixed multiple above the topcode point, with most researchers using a multiple between 1.3 and 1.5 (Autor, Katz, & Kearney, 2008; Juhn, Murphy, & Pierce, 1993; Lemieux, 2006). Implicit in this approach, regardless of the multiplier, is an assumption that the multiple is constant across years and across changes in the threshold level.

The multiples in this approach were partially derived from attempts to fit top earnings to a Pareto distribution. In particular, following the long-standing assumption that top earnings can be described by the Pareto distribution, numerous researchers have imputed the top of the earnings distribution based on those fit by a Pareto distribution (Bishop, Chiou, & Formby, 1994; Fichtenbaum & Shahidi, 1988; Heathcote, Perri, & Violante, 2010; Mishel, Bernstein, & Shierholz, 2013; Piketty & Saez, 2003; Schmitt, 2003).

The Pareto distribution is defined by the CDF:

$$P(X < x) = 1 - \left(\frac{x_c}{x}\right)^{\alpha} \quad (1)$$

Where: x is a given value of earnings (weakly) larger than x_c, x_c is the scale or cutoff parameter, and α is the shape parameter of the distribution. Since the Pareto distribution is scale-free, the mean above any threshold y is given as:

$$M(y) = \left(\frac{\alpha}{\alpha-1}\right) y \quad (2)$$

This provides a simple link to the fixed-multiple concept. By setting y as the topcode threshold, M(y) is the Pareto-imputed mean income above the threshold.

To use the Pareto distribution to estimate top earnings, one must first estimate the appropriate shape parameter. The most common approach is to assume that the distribution is

Pareto above some lower cutoff point (x_c) and choose a second cutoff point above that point—typically the topcode threshold itself (x_t) (Parker & Fenwick, 1983; Quandt, 1966; Shyrock & Siegel, 1975; Saez, 2000). The Pareto shape parameter is then:

$$\alpha = \frac{\ln\left(\frac{C}{T}\right)}{\ln\left(\frac{X_T}{X_C}\right)} \quad (3)$$

Where: *C* represents the number of individuals with earnings above the lower cutoff and *T* represents the number of individuals with earnings above the topcode threshold. Juhn, Murphy and Pierce (1993) report that their choice of cutoff points in the public-use CPS did not substantially impact their results. However, Schmitt (2003) using more recent public-use CPS data found that the choice of cutoff point could matter greatly, depending on the frequency of topcoding in the empirical distribution.

As we will illustrate below, this approach fails to provide reasonable estimates of top earnings in more recent public-use CPS data. This is partially because the earnings distribution may not be Pareto far enough below the public-use topcode threshold (if at all) to obtain reasonable estimates of the scale parameter. Additionally, it may be partially because using only two distribution points poorly measures the parameter.

We address the first of these concerns by estimating the shape of the Pareto distribution using the internal data with its less restrictive censoring. This allows us to reduce the portion of the distribution over which earnings must fit the Pareto distribution—1 to 2 percent rather than the 10 or 20 percent with the public-use CPS data (Mishel, Bernstein and Shierholz, 2013, for example, assume that 20 percent of the earnings distribution fits the Pareto distribution). To further improve the estimate, we use actual internal data when available for estimating top earnings, and only use the Pareto imputation for internally censored observations where the true value is unknown.

To address the second concern, we adapt an alternate, but rarely used, approach to estimating the Pareto scale parameter—applying a maximum likelihood formula to the empirical distribution. Polivka (2000) used this approach to analyze categorical weekly earnings data, but to our knowledge it has not been applied to continuous annual earnings data. Under this approach, the continuous, closed-form solution for estimating the Pareto parameter is:

$$\hat{\alpha} = \frac{M}{T \ln(X_T) + \sum_{x_m \le x_i < x_T} \ln(x_i) - (M+T)\ln(x_m)} \quad (4)$$

Where: M is the number of individuals with earnings between the lower cutoff and censoring point, T is the number of individuals with earnings at or over the topcode or censoring point, and x_i is the earnings of an individual. Using this formula allows individuals between the cutoff and censoring points to contribute to the PDF with their actual earnings, while those at the censoring point contribute to the CDF with the information that they have earnings at least as high as the censoring point.

In Figure 2 we compare the relative accuracy of the standard proportional and our maximum likelihood Pareto imputation approaches, along with the fixed-multiple approach from Lemieux (2006) and Katz and Murphy (1992) in capturing the top part of the earnings distribution censored in the public-use CPS. Since the Pareto cutoff point matters for both approaches, when using the public-use data we follow the approach of Mishel, Bernstein and Shierholz (2013) and assume the distribution is Pareto above the 80th percentile of the distribution.[3] Since we are using internal CPS data for the estimation using our Maximum Likelihood technique we can use a much higher cutoff, and assume the distribution is Pareto above the 99th percentile.[4]

[3] Alternate cutoffs of the 85th, 90th, and 95th percentiles were also considered. In general increasing the income cutoff for the lower bound of the estimation lowered the mean earnings of the top 5 percent.

[4] Alternate cutoffs of the 95th, 97^{th,} and 98th percentiles were also considered and produced largely consistent results for the mean earnings of the top 5 percent.

To compare the accuracy of the various series, we compare the mean annual earnings of the top 5 percent of the distribution for each with those in the Larrimore et al. (2008) cell-mean series based on the internal CPS data. The Larrimore et al. (2008) cell-mean series uses the internal CPS data to provide the mean value for each source of income for any individual whose income from that source is topcoded. But it is not designed to correct for internal censoring, and it treats each source of income at or above the internal censoring point as if it were equal to the censoring point. As a result, it is consistent with the Census Bureau's official income statistics, but both Larrimore et al. (2008) and the official Census Bureau statistics are known to represent an underestimate of the true top earnings of the population.

While the top earnings using the Pareto imputation based on public-use data and those using the fixed multiple series each slightly exceeds the top earnings from the Larrimore et al. (2008) cell-mean series in early years, neither does so after 1993 when improvements in Census Bureau collection procedures greatly improved the reporting of earnings by top earners. (See Jones & Weinberg, 2000 and Ryscavage, 1995 for details on this change.) Since the cell-mean series is a lower bound for top earnings, it is apparent that these previous efforts to capture the top part of the earnings distribution based solely on public-use CPS data understate their level at the upper tail since at least 1993.

In contrast to these earlier techniques, our Maximum Likelihood Pareto estimation of internally censored observations, in conjunction with the internal data when available, produces mean earnings of the top 5 percent which exceed those of Larrimore et al. (2008). In years before 1993, when the Census Bureau increased their internal censoring thresholds, this increased the mean earnings of the top 5 percent by between 7 and 14 percent in each year. In

more recent years, the gap has been smaller, ranging from a 1 to 6 percent increase over the values from Larrimore et al. (2008). [5]

Recognizing that these improved estimates are based on not generally available internal data, we created an enhanced cell-mean series which uses the actual internal data when available and these Pareto estimates for the internally censored data. This series, available in Appendix Table 1, allows researchers with only public-use data to obtain the best available estimate of top earnings in the CPS data.

IV: Comparison to Social Security Administration Records

Kopczuk, Saez, and Song (2010) provide the first research using administrative records data to analyze long-run earnings inequality. Their study uses Social Security Administration (SSA) earnings data from 1937 to 2004 to examine earnings inequality of Commerce and Industry workers between the ages of 18 and 70 with wages over $2,575 in 2004.[6] This study is the current gold standard of annual earnings inequality trends and hence an excellent benchmark for testing the validity of our CPS-based results. If results from Kopczuk, Saez, and Song (2010) can be replicated in the CPS data, then it validates the use of CPS data for analyzing earnings trends. To this end, we limit our data sample to Commerce and Industry workers and compare Gini coefficient results across the two datasets.

The earnings Gini for Commerce and Industry workers from Kopczuk, Saez, and Song (2010) are compared in Figure 3 to each of the topcode correction methods in Figure 2. While we do not have access to internal CPS data before 1967, to extend the comparison we go back to

[5] As a further test of the validity of the Pareto at this income level, we compare the Pareto scale parameter for the 95th, 97th, 98th, and 99th percentile. The Pareto parameters are generally stable, with the average difference between the maximum and minimum scale parameter in this range being just 16 percent apart. Pareto scale parameters are available upon request of the authors.
[6] Commerce and Industry workers are all non-farm, non-self-employment wage and salary workers not working in agriculture, forestry, fishing, hospitals, educational services, social services, religious organizations, private households, and public administration.

1963 using public-use CPS data.[7] Over these years topcoding was rare so no additional topcode corrections were required.[8] To match the data range in Kopczuk, Saez, and Song (2010) our comparison ends in 2004.

Since both the public-use data with no cell means and the series with a fixed multiplier of the topcode threshold miss the rise in top earnings, each of these series produce Gini coefficients that are well below the level of earnings inequality observed by Kopczuk, Saez, and Song (2010). The internal cell-means series from Larrimore et al. (2008) and the public-use Pareto are closer, but still understate inequality in almost all years since the early 1980s. In contrast, our enhanced cell-mean series, using the internal CPS data with Pareto estimates of internally censored observations, can largely match the trends from Kopczuk, Saez, and Song (2010) back to 1963. This provides evidence that, with appropriate corrections to capture the top of the earnings distribution, the public-use CPS data can be used to accurately measure and analyze United States earnings trends.

However, the inequality trends found here, and potentially those found by Kopczuk, Saez, and Song (2010), are sensitive to the decision to limit the sample to Commerce and Industry workers and to impose age and earnings restrictions. Figure 4 compares the Gini coefficient using the enhanced cell-mean series for two samples. The first is the Commerce and Industry worker sample, which matches the sample Kopczuk, Saez, and Song (2010) use, which we showed in Figure 3. The second is all workers with positive wage or salary income, regardless of industry or age. In the restricted sample, earnings inequality increases by 24.0 percent—0.375 to 0.465—from 1963 to 2004. In the full sample it increases by 5.4 percent—

[7] CPS data from 1961 is also available, however, the survey format changed between 1961 and 1963 which make the data incomparable between these years. Hence, we start our series in 1963, which is the earliest year for which we can create a consistent CPS series.

[8] No more than one worker was topcoded on wage and salary earnings each year over this period.

0.466 to 0.491. Hence, earnings inequality levels for the full population may be greater than Kopczuk, Saez, and Song (2010) observed in their subsample of Commerce and Industry workers but it may not have increased as much since the early 1960s.

V: Conclusion

Despite the common use of public-use CPS data for earnings inequality research, the previous methods of correcting for topcoding in the CPS data result in clear and substantial understatements of top earnings. In particular, both the fixed-multiple approach and Pareto estimates based solely on public-use CPS data understate the level of top earnings in the internal CPS data—which is also subject to censoring and thus represents a lower bound. Using a hybrid approach of internal data and Pareto imputations, this paper provides improved estimates of top earnings in the CPS data. Using this hybrid approach for estimating top earnings, we have produced an enhanced cell-mean series for use with the public-use data, which more closely approximates the actual level of top earnings in the population than was previously available in CPS data. When using the same sample restrictions used by Kopczuk, Saez, and Song (2010) with the public-use CPS data using our enhanced cell-mean series, we observe inequality levels that are consistent with those seen by Kopczuk, Saez, and Song (2010) for the subsample of U.S. workers in Commerce and Industry captured by administrative Social Security records. As a result, we believe that our series represents the best available measure of estimating top earnings in the CPS data and demonstrates that the CPS data can provide reasonable estimates of U.S. labor earnings trends.

References

Acemoglu, D. (2002). Technical Change, Inequality, and the Labor Market. *Journal of Economic Literature, 40* (1): 7-72.

Autor, D. H., Katz, L. F., & Kearney, M. S. (2008). Trends in U.S. Wage Inequality: Revising the Revisionists. *Review of Economics and Statistics, 90* (2), 300-323.

Bishop, J. A., Chiou, J.-R., & Formby, J. P. (1994). Truncation Bias and the Ordinal Evaluation of Income Inequality. *Journal of Business and Economic Statistics, 12*, 123-127.

Card, D., & DiNardo, J. E. (2002). Skill-Biased Technological Change and Rising Wage Inequality: Some Problems and Puzzles. *Journal of Labor Economics, 20* (4), 733-782.

Fichtenbaum, R., & Shahidi, H. (1988). Truncation Bias and the Measurement of Income Inequality. *Journal of Business and Economic Statistics, 6*, 335-337.

Heathcote, J., Perri, F., & Violante, G. L. (2010). Unequal We Stand: An Empirical Analysis of Economic Inequality in the United States: 1967-2006. *Review of Economic Dynamics, 13* (1), 15-51.

Jones, A. F., & Weinberg, D. H. (2000). *The Changing Shape of the Nation's Income Distribution.* Washington, DC: U.S. Census Bureau.

Juhn, C., Murphy, K. M., & Pierce, B. (1993). Wage Inequality and the Rise in Returns to Skill. *Journal of Political Economy, 101* (3), 410-442.

Katz, L. F., & Murphy, K. M. (1992). Changes in Relative Wages, 1963-87: Supply and Demand Factors. *Quarterly Journal of Economics, 107*, 35-78.

Kopczuk, W., Saez, E., & Song, J. (2010). Earnings Inequality and Mobility in the United States: Evidence from Social Security Data since 1937. *Quarterly Journal of Economics, 125*, 91-128.

Larrimore, J., Burkhauser, R. V., Feng, S., & Zayatz, L. (2008). Consistent Cell Means for Topcoded Incomes in the Public Use March CPS (1976-2007). *Journal of Economic and Social Measurement, 33* (2-3), 89-128.

Lemieux, T. (2006). Increased Residual Wage Inequality: Composition Effects, Noisy Data, or Rising Demand for Skill. *American Economic Review, 96* (2), 461-498.

Mishel, L., Bernstein, J., & Shierholz, H. (2013). *The State of Working America 12th Edition.* Ithaca, NY: Cornell University Press.

Parker, R., & Fenwick, R. (1983). The Pareto Curve and Its Utility for Open-Ended Income Distributions in Survey Research. *Social Forces, 61*, 872-885.

Piketty, T., & Saez, E. (2003). Income Inequality in the United States, 1913–1998. *Quarterly Journal of Economics, 118* (1), 1-39.

Polivka, A. (2000). Using Earnings Data from the Monthly Current Population Survey. *Unpublished Manuscript* .

Quandt, R. (1966). Old and New Methods of Estimation and the Pareto Distribution. *Metrika, 10,* 55-82.

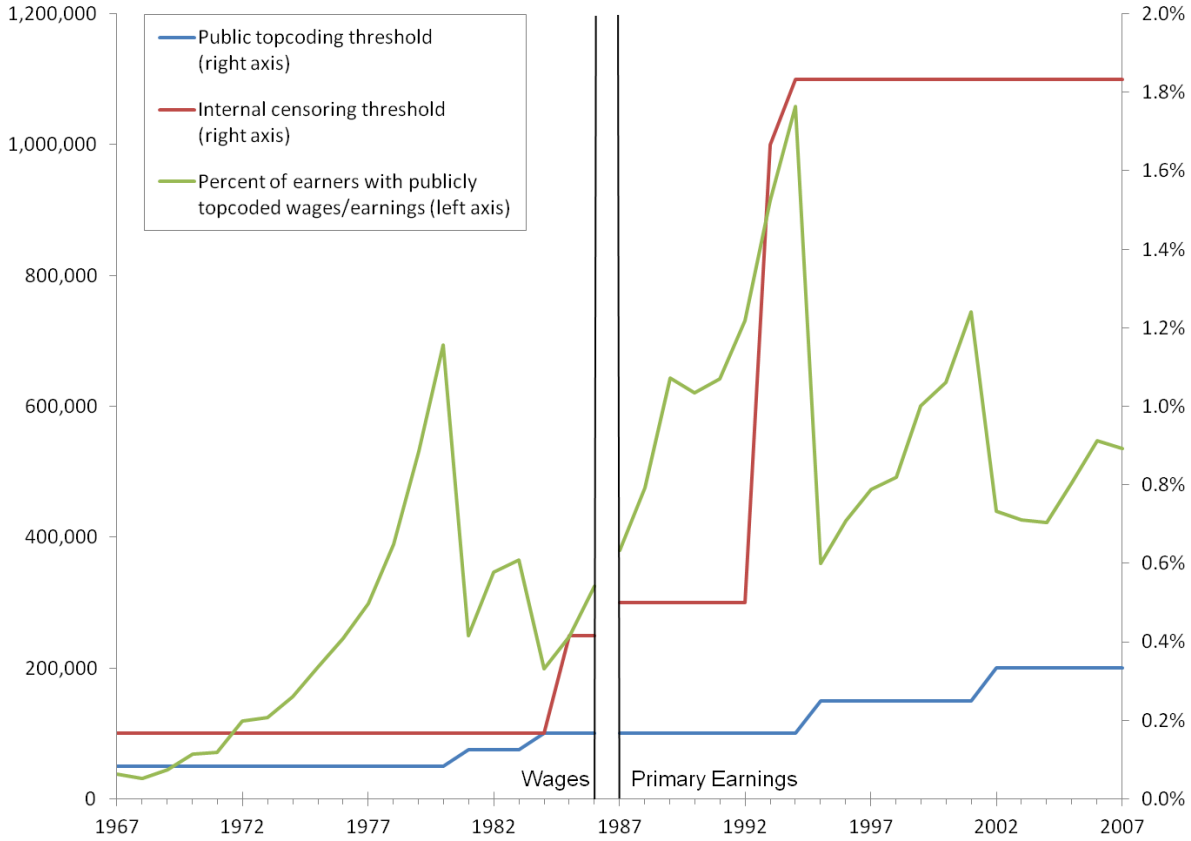
Ryscavage, P. (1995). A Surge in Growing Income Inequality? *Monthly Labor Review, 118* (8), 51-61.

Saez, E. (2000). Using Elasticities to Derive Optimal Income Tax Rates. *Review of Economic Studies, 68,* 205-229.

Schmitt, J. (2003). *Creating a Consistent Hourly Wage Series from the Current Population Survey's Outgoing Rotation Group, 1979-2002.* Washington, DC: Center for Economic and Policy Research.

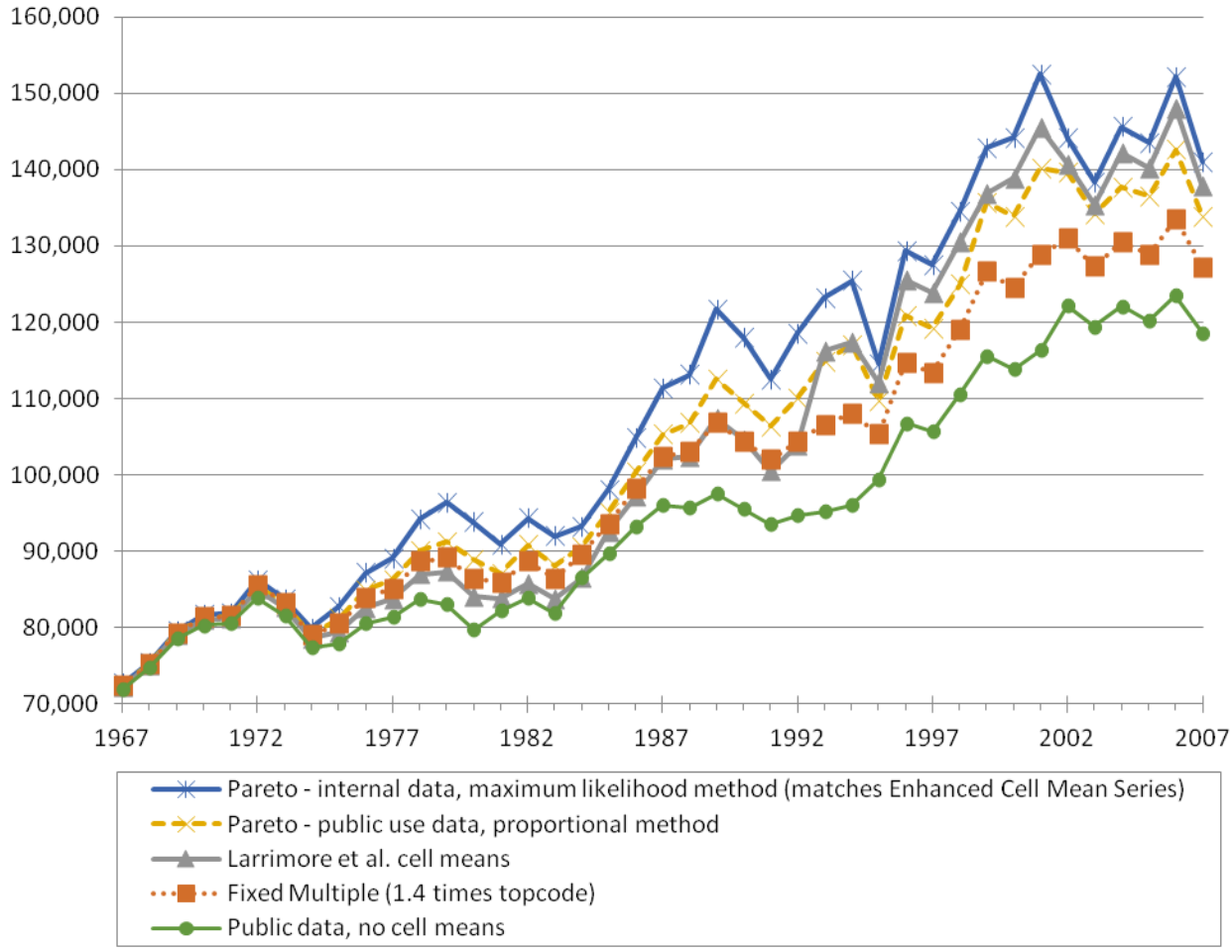
Shyrock, H., & Siegel, H. (1975). *The Methods and Materials of Demography.* Washington, DC: U.S. Government Printing Office.

Figure 1: Earnings topcodes and censoring thresholds in the March CPS data (1967-2007)



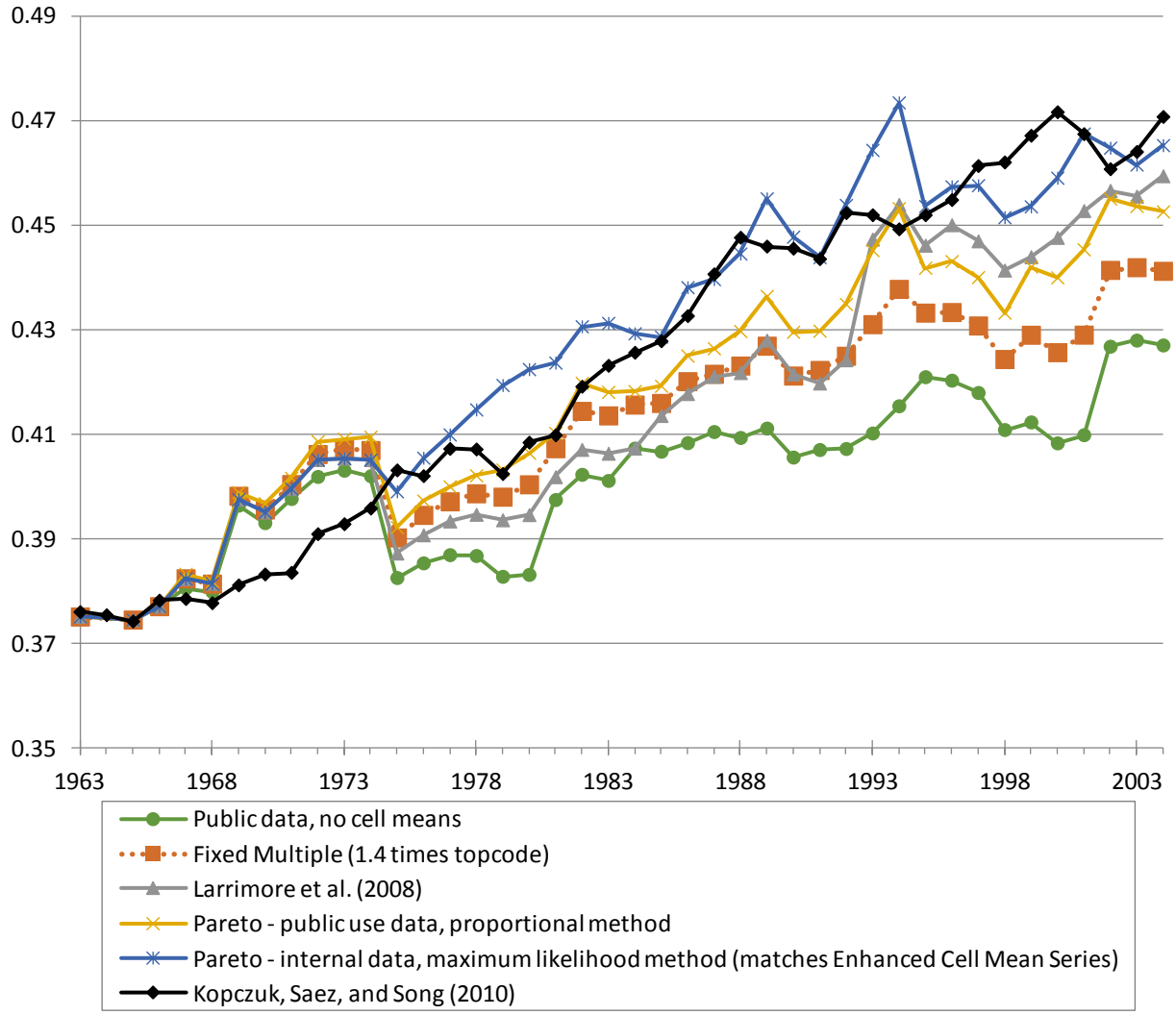
Sources: Author's calculation using Public Use and Internal March CPS Data.

Figure 2: Mean earnings of the top 5 percent of earners by topcode correction method (in 2010 dollars)



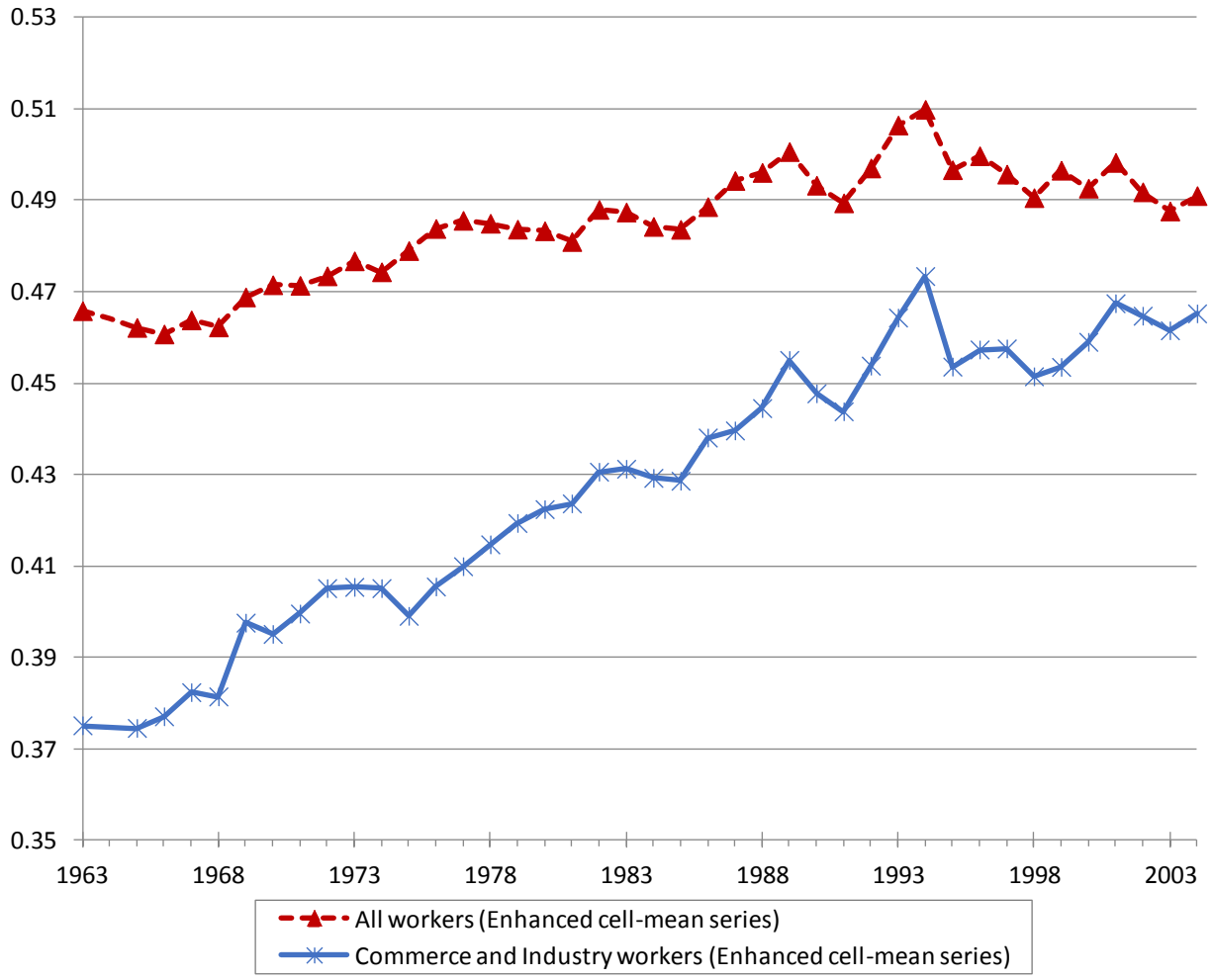
Sources: Author's calculation using Public Use and Internal March CPS Data.

Figure 3: Gini Coefficients for Commerce and Industry workers by topcode correction method, compared to Kopczuk, Saez, and Song (2010) estimates from SSA administrative records



Sources: Kopczuk, Saez, and Song (2010); Author's calculation using Public Use and Internal March CPS Data.
Notes: 1964 is missing in the CPS-based data series since there was no CPS survey in that year.

Figure 4: Gini Coefficients for all workers compared to Commerce and Industry workers, using internal CPS data with the Pareto correction method.



Source and Notes: See Figure 3.

Appendix Table 1: Enhanced Cell Means for wage and salary earnings (1967-1986) and for primary earnings (1987-2007)

Income Year	Mean Wage and Salary Earnings above Public Topcode	Income Year	Mean Primary Earnings above Public Topcode
1967	68,718.88	1987	236,346.60
1968	67,672.02	1988	232,933.60
1969	70,602.84	1989	246,791.60
1970	72,338.20	1990	241,900.10
1971	69,964.24	1991	225,628.70
1972	72,067.52	1992	238,452.00
1973	72,276.09	1993	238,452.00
1974	69,694.40	1994	238,936.10
1975	104,623.10	1995	357,275.70
1976	105,819.20	1996	372,871.20
1977	107,545.80	1997	393,602.50
1978	110,339.50	1998	387,119.90
1979	112,330.50	1999	343,966.10
1980	109,203.50	2000	383,150.80
1981	177,926.10	2001	392,626.30
1982	165,017.60	2002	471,696.00
1983	168,683.20	2003	449,855.00
1984	235,056.30	2004	481,784.40
1985	221,293.80	2005	472,174.80
1986	230,533.60	2006	490,588.50
		2007	459,918.10

Source: Author's calculation using Internal March CPS Data.
Note: Figures based on authors' calculation using internal CPS data and maximum likelihood Pareto fit at the 99th percentile of the earnings distribution. Income year records income in the year prior to the year of the March CPS survey. Enhanced cell means were not calculated for years before 1967 due to the lack of topcoding on earnings, when one individual or fewer was topcoded each year.