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

The research in this paper was conducted, in part, 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. All opinions are those of the authors and should not be attributed to the Census Bureau, the Federal Reserve Board, the Federal Reserve Banks, or their staff. We thank Stephen Jenkins and participants at the Census Bureau's Center for Economic Studies Research Conference for their helpful comments on earlier drafts of this paper. 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 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, Revised July 2015
JEL No. C81,D31,J01,J31

ABSTRACT

Inconsistent censoring in the public-use March CPS limits its usefulness in measuring labor earnings trends, as previous approaches for imputing topcoded earnings systematically understate top earnings. Using Pareto estimation methods with less-censored internal data, we create an enhanced cell-mean series to capture top earnings in the public-use data. Annual earnings inequality trends since 1963 using our series largely mirror those found by Kopczuk, Saez, and Song (2010) using Social Security Administration data for Commerce and Industry workers. When we extend our analysis to 2013 and consider all workers, earnings inequality levels are higher but its growth is more modest.

Philip Armour
RAND Corporation
1776 Main Street
Santa Monica, CA 90401-3208
parmour@rand.org

Jeff Larrimore
Federal Reserve Board
20th and Constitution Ave
Washington, DC 20551
jeff.larrimore@frb.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

I. INTRODUCTION

The public-use March Current Population Survey (CPS) is the primary source of data for tracking levels and trends in U.S. labor earnings and labor earnings inequality, and for explaining their causes. This literature has especially focused on changing returns to education in the labor market as well as whether the rise in wage and earnings inequality in the 1980s was part of a long-run secular trend or an episodic event (Autor 2015; Autor, Katz, and Kearney 2008; Card and DiNardo 2002; Goldin and Katz 2007; Hubbard 2011; Juhn, Murphy, and 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, and Kearney 2008; Katz and Murphy 1992; Juhn, Murphy, and 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

[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. Similar techniques to those used in the March CPS data have been employed in the May ORG sample to correct for topcoding, including replacing topcoded earnings with a fixed multiple of the topcode threshold (see e.g. Acemoglu and Autor 2010).

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 data understate mean earnings at the top of the earnings distribution and hence also understate 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 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 have access only 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. Having done so, we then extend our analysis to 2013 and consider all workers. While earnings inequality levels are higher when considering all workers rather than just Commerce and Industry workers, its growth is more modest.

II. DATA

The March CPS survey contains a comprehensive set of questions on sources of household earnings, including labor earnings which are the focus of this study.[2] These data are

[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.

collected annually by the Census Bureau and the CPS is one of the primary sources of data for research on income and earnings trends in the United States (see e.g. Autor, Katz, and Kearney 2008; Burkhauser et al. 2012; Card and DiNardo 2002; Gottschalk and Danziger 2005).

A known limitation of the March CPS data is that incomes are topcoded in the public-use data and censored at higher thresholds in the internal data. These topcoding and censoring thresholds change on an ad-hoc basis. Figure 1 provides an overview of these changes for annual wage earnings from 1967-1986 and for primary labor earnings, which are primarily wages, from 1987-2013.[3] Internal topcoding thresholds, with the exception of 1984, have always been higher than those in the public-use data but became substantially so after 1984. As a result while the number of individuals topcoded in the internal data has risen somewhat since then, the number topcoded in the public-use data has risen much more. Figure 1 shows this growth, as measured by percent of individuals with earnings above the public topcode (right axis), is erratic, rising when the Census Bureau holds topcodes nominally constant and quickly falling when they raise the topcodes.

We use both the public-use and internal CPS data to illustrate the impact of different correction techniques for topcoded earnings on earnings trends. Our preferred technique is derived from the internal CPS data, but researchers without access to the internal data can use it with the public version of the CPS data.

III: ESTIMATING TOP EARNINGS

[3] 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.

Most researchers measuring long-term trends in earnings with public-use CPS data use 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, and Kearney 2008; Juhn, Murphy, and 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 are 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 impute the top of the earnings distribution based on those fit by a Pareto distribution (Bishop, Chiou, and Formby 1994; Fichtenbaum and Shahidi 1988; Heathcote, Perri, and Violante 2010; Hubbard 2011; Mishel, Bernstein, and Shierholz 2013; Piketty and Saez 2003; Schmitt 2003).

The Pareto distribution is defined by the cumulative distribution function (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 and Fenwick 1983; Quandt 1966; Shyrock and 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 and because using only two distribution points may poorly measure 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.[4] This allows us to reduce the portion of the distribution over which earnings must fit the Pareto distribution—1 to 2% rather than the 10 or 20% with the public-use CPS data (Mishel, Bernstein, and Shierholz 2013, for example, assume that 20% of the earnings distribution fits the Pareto distribution).

[4] Our access to internal CPS data extends through 2007. However in 2010 the Census Bureau began providing "rank proximity swapped" incomes in the public-use data for topcoded incomes, and did so for earlier years retroactively. This approach is intended to yield the same distribution for each earnings source as in the internal data, but randomly swaps earnings values among topcoded individuals to protect confidentiality. Using these data with the same Pareto imputation technique for observations above the internal censoring point, we extend our series to include years after 2007. To ensure consistency, we replicate the enhanced cell-mean series using internal data using the rank-proximity swapped data and find consistent results. Results of this consistency check are available upon request from the authors.

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. We modify the widely used Maximum Likelihood Hill estimator(Hill, 1975) such that earners under the topcode contribute an observation indicating their reported earnings to the likelihood function, while earners above the topcode contribute an observation indicating that they earn at least the topcoded amount. This differential contribution provides an unbiased estimate of the Pareto parameter while utilizing all available information, and it has been used to fit other distributions, such as the Generalized Beta of the Second Kind, to topcoded earnings data (Jenkins et al. 2011). While Polivka (2000) uses this modified Hill estimator to analyze categorical weekly earnings data, 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 above 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 CDF 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.

To further improve the estimate of top earnings, rather than imputing all values the Census Bureau censors in the public-use data, 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. In order to facilitate the use of these better estimates of top

incomes, which combine actual internal data with imputations of censored observations, we create an enhanced cell-mean series consisting of the mean earnings of publicly topcoded individuals based on our combined set of internal data and Pareto estimates of censored observations. Researchers can use this series, available in Appendix Table 1, in conjunction with the public-use March CPS to obtain the best available estimate of top earnings based on these publically available data.

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.[5] 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.[6]

To compare the accuracy of the various series, we compare the mean annual earnings of the top 5% 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

[5] We also used cutoffs at the 85th, 90th, and 95th percentiles. In general increasing the income cutoff for the lower bound of the estimation lowered the estimated mean earnings of the top 5%.
[6] We also used cutoffs at the 95th, 97th, and 98th percentiles and produced largely consistent results for the mean earnings of the top 5%.

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 exceed the top earnings from the Larrimore et al. (2008) cell-mean series in early years, neither does so after 1993 when changes in Census Bureau collection procedures greatly improved the reporting of earnings by top earners. (See Jones and Weinberg 2000 and Ryscavage 1995 for details on this change.) Since the cell-mean series is a lower bound for top earnings, it is clear 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% which exceed those of Larrimore et al. (2008). In years before 1993, the impact of this adjustment are small. However, in more recent years, the addition of an imputation of censored earnings increases the average earnings of the top 5% by as much as 10% over the values from Larrimore et al. (2008).[7]

In comparing the series it may appear counterintuitive that imputing censored earnings using a Pareto distribution increased top earnings by *more* after 1993 relative to the Larrimore et al. (2008) cell mean series, when the Census Bureau increased their censoring threshold, than it did prior to that year. However, in addition to increasing the censoring threshold, the Census Bureau implemented other survey design changes in that year—such as electronic data

[7] 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% apart. Pareto scale parameters are available upon request of the authors.

collection—that fundamentally changed the shape of the upper tail of the observed income distribution. This can be observed in Table 1, which shows the percent of respondents in the data with earnings in high-income ranges (not adjusted for inflation). In each year from 1990 through 1992, between 0.18 and 0.24% of respondents reported earnings of $199,999 or more—and 0.06% of respondents reported income at or above the $299,999 internal censoring threshold for those years. In 1993, the fraction of respondents reporting an income of at least $199,999 nearly doubled to 0.43% of respondents. Similarly, the fraction with incomes of at least $299,999 tripled to 0.19% of respondents.

If the top of the income distribution does follow the Pareto distribution, this increase in the number of individuals with earnings near the censoring threshold suggests that there is also a longer right-tail of earnings above the threshold. Thus, the improvements in data collection in 1993 increased information about both the observed and unobserved portions of the distribution. The results in Figure 2, where we use the internal data with a Pareto imputation for censored values, demonstrate that the break in the data series in the raw internal data (Jones and Weinberg 2000 and Ryscavage 1995) may, in fact, *underestimate* the improvements in capturing top incomes occurring in that year. Recognizing that this trend break is the result of new collection procedures and not changes to topcoding, we correct for the break using the standard approach from Atkinson, Piketty, and Saez (2011) and Burkhauser et al. (2012) and upwardly adjust inequality measures from all years before 1993, thus assuming no inequality change in the 1992-1993 trend-break year.

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 25 and 60 with wages over $2,575 in 2004, indexed by nominal average wage growth for earlier years.[8] This minimum earnings restriction represents one-fourth of the earnings an individual working full time for a year (2,000 hours) at the federal minimum wage would receive each year. 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 impose the same age and minimum earnings restriction so that we can compare Gini coefficient results across the two datasets.

In Figure 3 we compare the earnings Gini for this subsample of workers from Kopczuk, Saez, and Song (2010) to our Pareto-adjusted income series as well as to estimates using the Larrimore et al. (2008) series, which were previously the best estimates of top earnings in the CPS data.[9] While we do not have access to internal CPS data before 1967, to extend the comparison we go back to 1963 using public-use CPS data.[10] Over these earlier years from 1963-1966, topcoding was so rare that no additional topcode corrections are required.[11]

[8] 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.

[9] While the cell means procedures used both by Larrimore et al. (2008) and in our enhanced cell mean series removes the variance of topcoded earnings, Larrimore et al. demonstrated that the fraction of respondents who are topcoded is small enough that obtaining the level of their income is sufficient to correct for the trend in the Gini coefficient even without their variance.

[10] CPS data from 1961 are 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.

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

Between 1967 and 1994, the inequality trend between the CPS data with our Pareto correction and the Kopczuk, Saez, and Song (2010) series using Social Security records is remarkably similar. In 1995, top earnings in the CPS series falls, resulting in a level of inequality that is approximately two-Gini points below the Kopczuk, Saez, and Song series. However, after that divergence the inequality trend between 1995 and 2004 continues to grow at a similar pace across the two series. Despite this divergence, our new series using the Pareto correction more closely matches the estimates from Kopczuk, Saez, and Song (2010) than does the Larrimore et al. cell mean series.

This provides evidence that our correction improves the ability of the public-use CPS data to accurately measure and analyze U.S. earnings levels and trends.

V: IMPACT OF AGE AND EARNINGS RESTRICTIONS

After largely matching the earnings inequality trends from Kopczuk, Saez, and Song (2010), we now focus on the extent to which limiting the sample to Commerce and Industry workers and imposing age and earnings restrictions influences observed inequality trends. In Figure 4, while still excluding labor earnings from self-employment, we compare the Gini coefficient for labor earnings we get using our enhanced cell-mean series for all workers with any earnings to the Gini coefficient for labor earnings we got using the sample restrictions imposed by Kopczuk, Saez, and Song in Figure 3. In the restricted sample, earnings inequality increases by 16.7%—0.378 to 0.441—from 1963 to 2013. When looking at workers in all industries, the level of inequality was similar, but the growth slowed from 16.7% to 11.1%. When we remove the restriction of considering only workers aged 25 to 60, and consider workers of all ages with earnings above the $2,575 minimum earnings restriction, the level of

inequality increases (in 2013 the Gini coefficient for our all-age group is 1.0% higher than the initial Commerce and Industry workers sample, 0.445 compared to 0.441), but its growth since 1963 is even slower. Without the age restriction, earnings inequality in 2013 was 5.2% above than in 1963.

Finally, we remove the $2,575 minimum earnings restriction and include all workers with earnings of at least $1 in the sample. Inequality, in 2013, in this fuller sample of workers is 10.2% higher than it is in the initial Commerce and Industry workers sample. But rather than increasing since 1963, earnings inequality is 2% lower in 2013 than it was in 1963. In contrast to the levels and trends in earnings inequality Kopczuk, Saez, and Song (2010) and we observe in their subsample of workers, in our full sample of workers we find the level of inequality is higher but its growth is less.

V: CONCLUSION

Inconsistent censoring in the public-use March Current Population Survey (CPS) limits its usefulness in measuring labor earnings levels and trends. We find that previous approaches for imputing topcoded earnings systematically understate 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. Our hybrid approach of internal data and Pareto imputations provides better estimates of top earnings in the CPS data. Using our hybrid approach, we create an enhanced cell-mean series for use with the public-use data that will allow researchers to more closely approximate the actual level of top earnings in CPS data. Using public-use CPS data together with our enhanced cell-mean series and mimicking Kopczuk, Saez, and Song (2010)

sample restrictions, we observe labor earnings inequality levels that are more consistent with those Kopczuk, Saez, and Song (2010) report 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. "Technical Change, Inequality, and the Labor Market." *Journal of Economic Literature,* 40(1), 2002, 7–72.

Acemoglu, D. and D. H. Autor. "Skills, Tasks and Technologies: Implications for Employment and Earnings," in *Handbook of Labor Economics*, Volume 4B, edited by O. Ashenfelter and D. Card, Amsterdam: Elsevier, 2010, 1043–172.

Atkinson, A. B., T. Piketty, and E. Saez. "Top Incomes in the Long Run of History." *Journal of Economic Literature*, 49(1), 2011, 3–71.

Autor, D. H., L. F. Katz, and M. S. Kearney. "Trends in U.S. Wage Inequality: Revising the Revisionists." *Review of Economics and Statistics,* 90(2), 2008, 300–23.

Bishop, J. A., J. R. Chiou, and J. P. Formby. "Truncation Bias and the Ordinal Evaluation of Income Inequality." *Journal of Business and Economic Statistics,* 12, 1994, 123–7.

Burkhauser, R.V., S. Feng, S. P. Jenkins, and J. Larrimore. "Recent Trends in Top Income Shares in the United States: Reconciling Estimates from March CPS and IRS Tax Return Data." *Review of Economics and Statistics*, 94(2), 2012, 371–88.

Card, D., and J. E. DiNardo. "Skill-Biased Technological Change and Rising Wage Inequality: Some Problems and Puzzles." *Journal of Labor Economics*, 20(4), 2002, 733–82.

Fichtenbaum, R., and H. Shahidi. "Truncation Bias and the Measurement of Income Inequality." *Journal of Business and Economic Statistics,* 6, 1988, 335–7.

Goldin, C. and L. F. Katz. "Long-Run Changes in the Wage Structure." *Brookings Papers on Economic Activity,* 2, 2007, 135–67.

Gottschalk, P., and S. Danziger. "Inequality of Wage Rates, Earnings and Family Income in the United States, 1975–2002." *Review of Income and Wealth*, 51(2), 2005, 231–54.

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

Hill, B. M. "A Simple General Approach to Inference About the Tail of a Distribution." *The Annals of Statistics,* 3(5), 1975, 1163–74.

Hubbard, W. H. J. "The Phantom Gender Difference in the College Wage Premium." *Journal of Human Resources,* 46(3), 2011, 568–86.

Jenkins, S. P., R. V. Burkhauser, S. Feng, and J. Larrimore. "Measuring Inequality Using Censored Data: A Multiple-Imputation Approach to Estimation and Inference." *Journal of the Royal Statistical Society: Series A*, 174(1), 2011, 63–81.

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

Juhn, C., K. M. Murphy, and B. Pierce. "Wage Inequality and the Rise in Returns to Skill." *Journal of Political Economy*, 101(3), 1993, 410–42.

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

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

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

Lemieux, T. "Increased Residual Wage Inequality: Composition Effects, Noisy Data, or Rising Demand for Skill." *American Economic Review*, 96(2), 2006, 461–98.

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

Parker, R., and R. Fenwick. "The Pareto Curve and Its Utility for Open-Ended Income Distributions in Survey Research." *Social Forces*, 61, 1983, 872–85.

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

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

Quandt, R. "Old and New Methods of Estimation and the Pareto Distribution." *Metrika*, 10, 1966, 55–82.

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

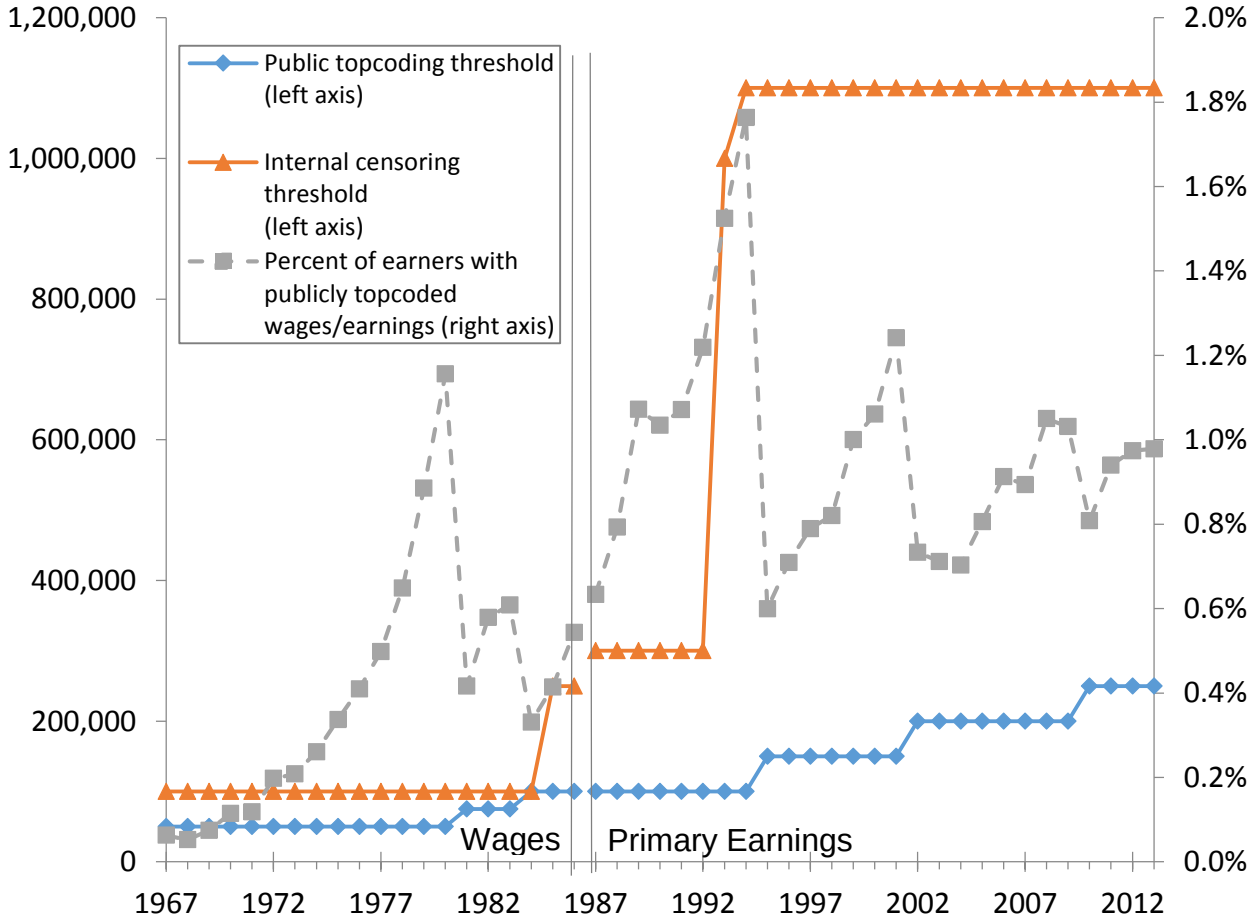
Saez, E. "Using Elasticities to Derive Optimal Income Tax Rates." *Review of Economic Studies*, 68, 2000, 205–29.

Schmitt, J. *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, 2003.

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

FIGURE 1

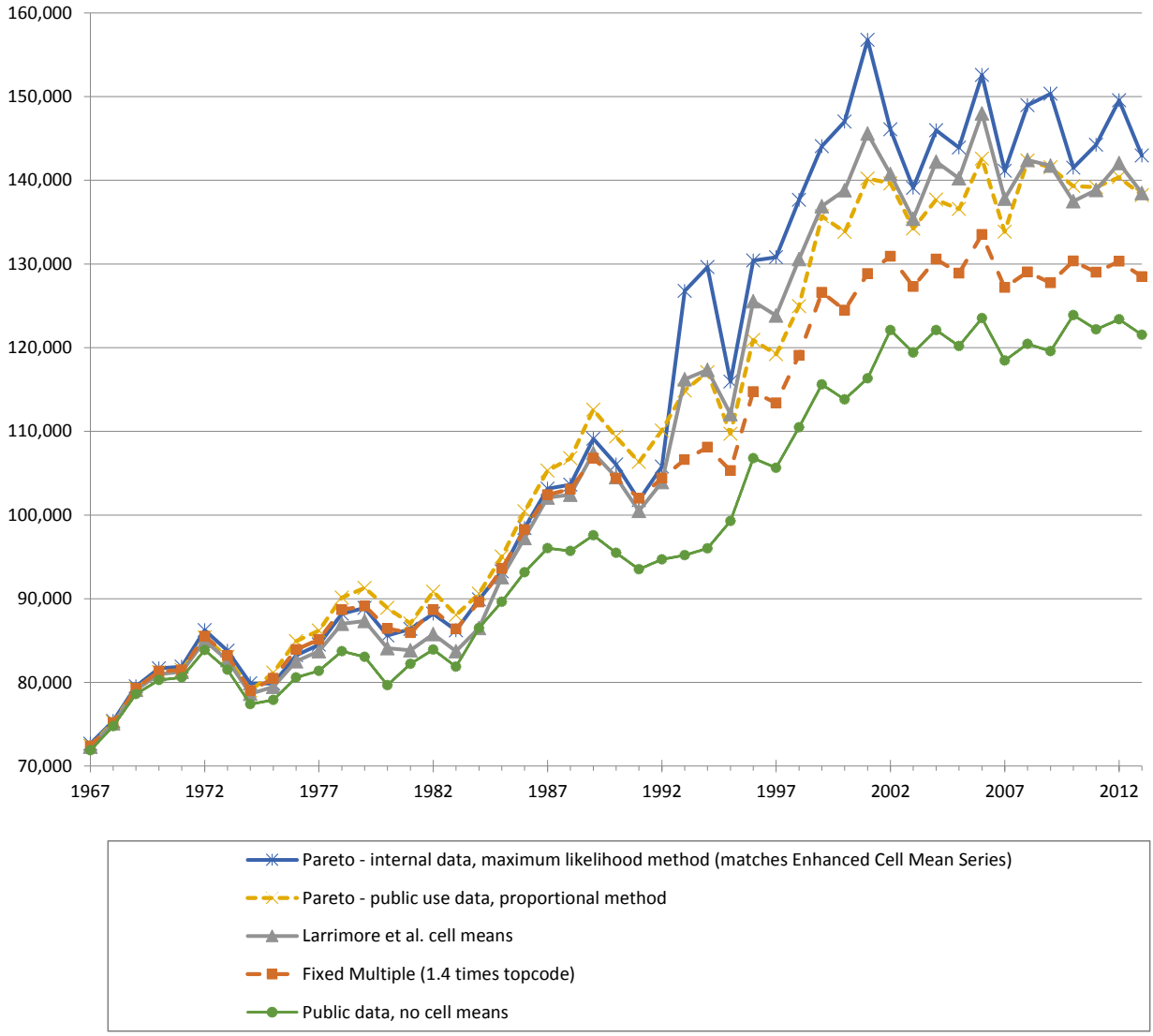
Earnings Topcodes and Censoring Thresholds in the March CPS Data (1967-2013)



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

FIGURE 2

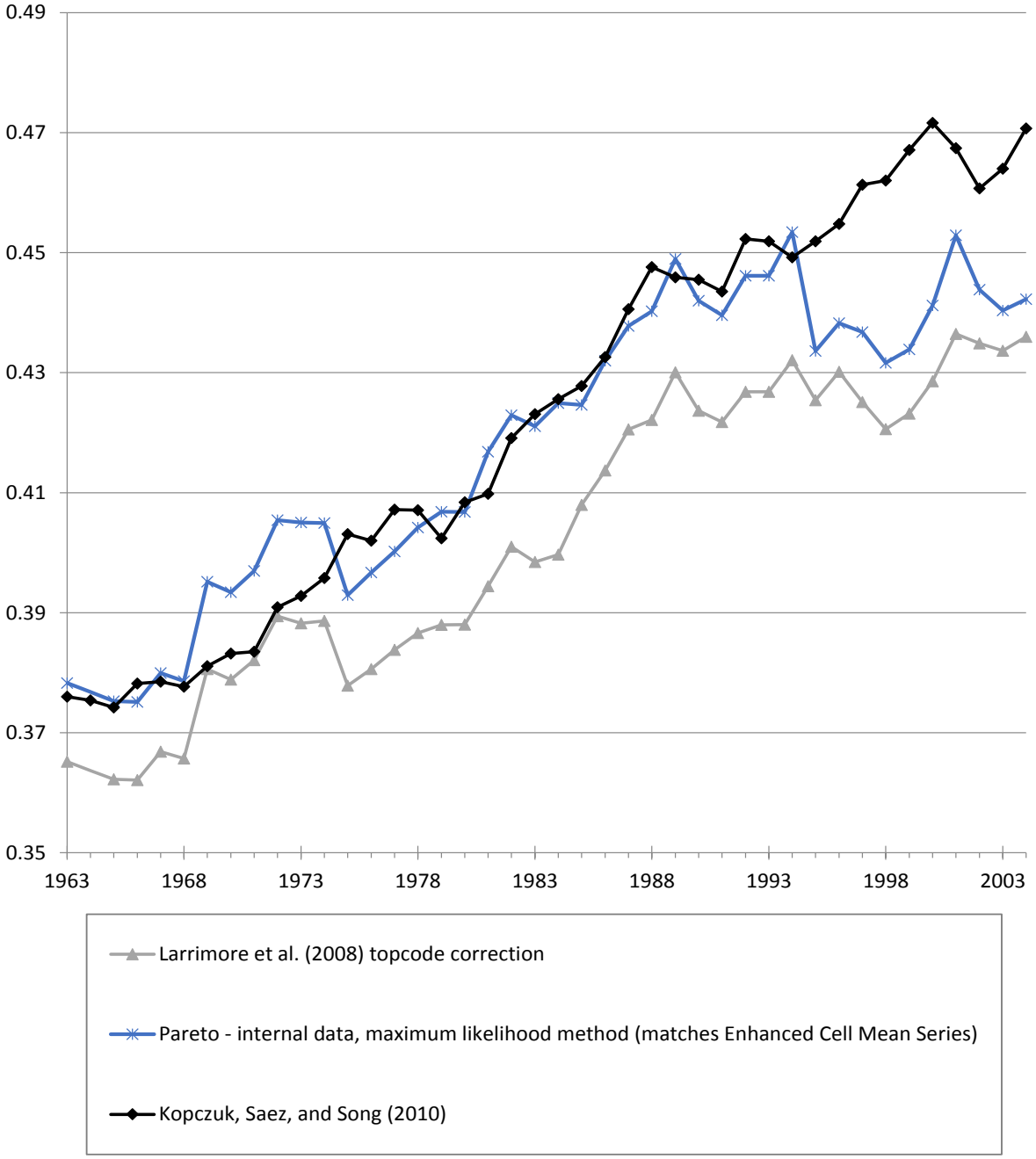
Mean Earnings of the Top 5% of Earners by Topcode Correction Method (in 2010 dollars)



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

FIGURE 3

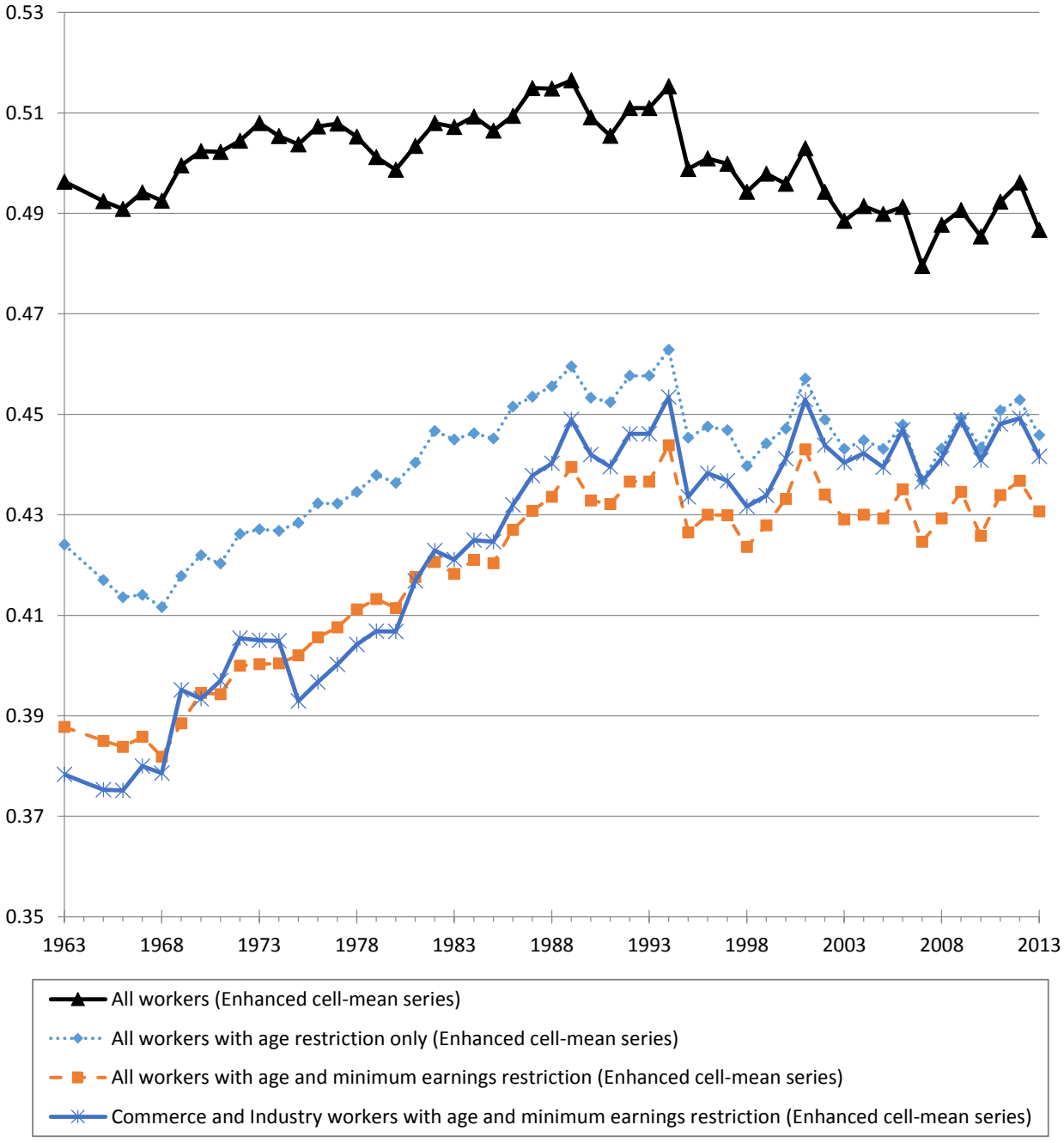
Gini Coefficients for Commerce and Industry Workers Using Pareto Correction, Compared to Kopczuk, Saez, and Song (2010) Estimates from SSA Administrative Records



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. All series are limited to Commerce and Industry workers aged 25-60 with at least $2,575 in 2004 earnings, indexed to nominal average wage growth in other years. The series end in 2004 to match the sample period from Kopczuk, Saez, and Song (2010). All CPS-based data prior to 1993 is corrected using the approach from Atkinson, Piketty, Saez (2011) and Burkhauser et al. (2012) described in text.

FIGURE 4

Gini Coefficients for All Workers Compared to Commerce and Industry Workers, Using Internal CPS Data with the Pareto Correction Method



Source: See Figure 3.
Notes: 1964 is missing in the CPS-based data series since there was no CPS survey in that year. Series with age restrictions are limited to individuals aged 25-60. Series with minimum earnings restrictions are limited to individuals with earnings of at least $2,575 in 2004 indexed by nominal average wage growth for other years.

TABLE 1

Percent of Earners in High Earnings Brackets (in Nominal Dollars) in Years Before and After Census's 1993 Changes to Data Collection Procedures

	Before changes to data collection procedures			After changes to data collection procedures		
High Earnings bracket	1990	1991	1992	1993	1994	1995
$199,999 to $299,998	0.16	0.12	0.18	0.24	0.26	0.24
$299,999 or above	0.06	0.06	0.06	0.19	0.20	0.21

Source: Public-use CPS data with rank proximity swap data.

APPENDIX TABLE 1

Enhanced Cell Means for Wage and Salary Earnings (1967-1986) and for Primary Earnings (1987-2013)

Income Year	Mean Wage and Salary Earnings above Public Topcode	Income Year	Mean Primary Earnings above Public Topcode
1967	68,718.88	1987	155,167.85
1968	67,672.02	1988	153,957.31
1969	70,602.84	1989	161,368.84
1970	72,338.20	1990	161,071.86
1971	69,964.24	1991	149,446.92
1972	72,067.52	1992	157,823.42
1973	72,276.09	1993	240,177.96
1974	69,694.40	1994	240,310.44
1975	68,484.37	1995	362,741.41
1976	69,622.58	1996	374,699.39
1977	70,377.94	1997	398,231.55
1978	72,473.37	1998	387,378.22
1979	77,877.91	1999	347,774.63
1980	76,067.81	2000	419,886.77
1981	116,517.60	2001	390,670.08
1982	108,677.47	2002	470,904.67
1983	110,527.71	2003	445,997.33
1984	152,540.90	2004	477,597.05
1985	147,726.89	2005	474,259.17
1986	151,170.99	2006	538,416.97
		2007	467,984.50
		2008	464,928.55
		2009	501,245.62
		2010	522,694.47
		2011	572,896.45
		2012	626,923.39
		2013	558,843.72

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. Enhanced cell means for years since 2008 are based off of the "rank proximity swap" data from the Census Bureau, which we observe produces comparable results to those based off of internal CPS data.