Chapter 4.
BETTER ESTIMATES OF INCOME AND ITS DISTRIBUTION IN THE PUBLIC-USE MARCH CURRENT POPULATION SURVEY[1]

RICHARD V. BURKHAUSER[2] AND
JEFF LARRIMORE
CORNELL UNIVERSITY

The March Current Population Survey (CPS) is the primary data source used by public policy researchers and administrators to investigate trends in U.S. income and its distribution. For confidentiality reasons, the U.S. Census Bureau topcodes each of the 24 sources of income (11 income sources prior to 1988) in the public-use CPS. However, this topcoding—the suppression of income values above some level in the public-use CPS data for confidentiality reasons—has not been consistent over time. Below we discuss a new set of papers that offer a solution to this problem using a series of created values, which, when used together with the public-use CPS data, will closely approximate income and inequality levels and their trends based on the internal CPS.

THE PROBLEM

Unsystematic topcoding in the public-use CPS data inconsistently restricts the ability of researchers to fully capture income and its distribution over time. Thus, researchers using scalar inequality measures such as the Gini coefficient with the public-use CPS will unsystematically understate both levels and trends in income inequality. In Figure 1, we compare the Gini coefficients using the topcoded public-use March CPS data with those obtained using the internal CPS data. The Public-Unadjusted series is the Gini calculated from the public-use March CPS data exactly as it is provided by the Census Bureau with no adjustments made for topcoding. Similarly, the Internal-Unadjusted series is the Gini calculated from the internal March CPS data with no adjustments made to the data.

[Place Figure 1 about here]

Prior to 1995, the Census Bureau reported topcoded incomes as equal to the topcode threshold, which reduced the observed income of all topcoded individuals. Using the topcode threshold led to a substantial understatement in the level of inequality and distorted the trends in inequality. Starting in 1995, the Census Bureau began providing cell means—the mean of all topcoded values for each income source—which, as can be seen in Figure 1, allows researchers using the Public-Unadjusted CPS data to closely match results from the internal CPS data after 1995.

Unfortunately, researchers interested in longer-term income trends have largely ignored this valuable information because cell means were not available for prior years. Instead, many researchers have simply recoded topcoded incomes as equal to the topcode threshold or have opted to use relatively simple measures of inequality, such as the 90/10 ratio, to attempt to avoid topcoding problems. However, Burkhauser, Feng, and Jenkins (2009) show that even researchers using inequality measures, such as the 90/10 ratio, run the risk of understating income inequality. This is because even individuals whose total household size-adjusted income is relatively modest and hence below the 90th percentile may have one or more topcoded sources of household income.

Here we discuss how newly available information that we created about the means and variances of top incomes from the internal CPS data can be used in conjunction with public-use CPS data to largely remedy the problems resulting from inconsistent topcoding. Using the extended cell mean series from Larrimore et al. (2008) and the variances of topcoded incomes from Burkhauser, Feng, and Larrimore (2008), it is now possible to better capture the top part of the income distribution with public-use CPS data and closely replicate results found using the internal CPS data used by the Census Bureau for producing their official inequality statistics (U.S. Census Bureau, various years).

USING CELL MEANS AND VARIANCES TO OBTAIN BETTER ESTIMATES OF TOP INCOMES

Despite the ability of cell means to closely replicate the results from internal CPS data, the lack of cell means prior to 1995 has dissuaded researchers from using them when looking at long-term trends using public-use CPS data. Thus, in Larrimore

[1] Support for this research from the National Science Foundation (award nos. SES-0427889, SES-0322902, and SES-0339191) and the National Institute for Disability and Rehabilitation Research (H133B040013 and H133B031111) are cordially acknowledged.

[2] Burkhauser completed this paper while he was the R. I. Downing Fellow in Social Economics in the Faculty of Economics and Commerce, University of Melbourne.

et al. (2008) we accessed internal CPS data and used it to calculate and distribute a cell mean series to the public going back to 1975. We also show that using this series will greatly improve the ability of researchers using only the public-use CPS to capture the top part of the income distribution.

A cell mean more accurately captures the level of unobserved income for a given source of income in the public-use CPS data but does not provide information about its distribution. Therefore, in Burkhauser, Feng, and Larrimore (2008), we also calculate and distribute to the research community the variances of topcoded incomes for each source of income in the internal CPS. By using information on both mean and variance, it is possible to impute different total household income values for each topcoded individual in the public-use CPS data. While doing so, in general, will not match an individual's actual income, it will allow the resulting distribution of income to more accurately match the distribution found in the internal CPS. Below we report how we have used both our cell mean and variance series to better estimate income and its distribution using public-use CPS data.

CREATING MORE CONSISTENT MEASURES OF TRENDS IN HOUSEHOLD INCOME INEQUALITY

Burkhauser et al. (2008) analyze levels and trends in inequality, using Gini coefficients between 1975 and 2004 derived from the internal CPS, and compare them with estimates from several series derived from the public-use CPS. The series and their sources (internal or public-use data) are described in Table 1.

[Place Table 1 about here]

Figure 2, taken from Burkhauser et al. (2008), shows that those who simply use the unadjusted public-use CPS (Public-Unadjusted) will find income inequality jumps dramatically between 1994 and 1995; i.e., the Gini value increases from 0.395 to 0.422, a single year change far greater than in any prior or subsequent year. An increase in the topcode threshold as well as the use of Census Bureau derived cell mean values for all values above the topcode threshold caused this jump. Using the unadjusted internal CPS (Internal-Unadjusted), we find no such increase between 1994 and 1995. Rather, what is happening is that prior to 1995, the Public-Unadjusted CPS Gini values substantially understate income inequality because they fail to fully account for income values above the topcodes. Once the Census Bureau provided the mean value of all these topcoded values, the now more precise Public-Unadjusted CPS Gini values match the higher Internal-Unadjusted Gini values. Failure to account for this change in methodology will grossly overstate U.S. inequality increases before and after 1994–1995.

[Place Figure 2 about here.]

Simply ignoring Census Bureau mean values after 1994, as can be seen in the Public-NoMean Gini series, will not solve this problem. This series still inconsistently topcodes high values, and it underestimates inequality after 1994, as can be seen by the way its Gini values fall further and further below the Internal-Unadjusted Gini values. In contrast, when we use the public-use CPS together with our extended mean series (Public-Mean) in Figure 2, we come much closer to matching the Internal-Unadjusted Gini values in every year.

However, the internal CPS data is itself censored, albeit to a substantially smaller extent than the public-use CPS. Hence, it too has time-inconsistencies, especially in 1992–1993, as can be seen by the jump in the Internal-Unadjusted Gini values between these years. To both control for inconsistent censoring and capture the missing part of the internal CPS data, we use a multiple imputation approach in which, for each year, out-of-sample values for topcoded observations in the internal CPS are imputed based on lower in-sample values that we do have. Unsurprisingly, as can also be seen in Figure 2, we find that compared to estimates derived from our multiple imputation approach (Internal-Adjusted), all the other series understate the level of inequality in all years.

However, just as was the case for our Public-Mean series and the Internal-Unadjusted series it replicated, the Internal-Adjusted series reveals the same trends—an increase in inequality over the entire period 1975–2004 but with a rate of increase noticeably lower after 1993 compared to before 1993. In each series, average inequality increases much more prior to 1992 than after 1993. In addition, in each series, the jump in 1992–1993 is far higher than in any other period. This is consistent with the argument that a change in the measurement of inequality rather than a real change in inequality is its cause.

To further test our Internal-Adjusted series, we compared our results to those derived by Piketty and Saez (2003) using Internal Revenue Service (IRS) administrative files. We did so by comparing our estimates of the share of income held by the wealthiest 10 percent of the population to that found by Piketty and Saez (2003). As can be seen in Figure 3, we find that these two estimates of the income share held by the top 90th–95th percentiles and the 95th–99th percentiles have remarkably similar levels and trends, especially given that our income units differ as we are observing household income while they are observing the adjusted gross income of tax units. It is only in the share of income held by the richest 1 percent where our levels and trends differ. Piketty and Saez (2003) not only find that a larger share of income is held by this group but that the growth in their share has been greater over time. Differences in the measures of income we use explain differences in levels. However, it is unclear whether the differences in trends are due to an increasing inability of even the internal CPS to capture the very top part of the income distribution or to behavioral changes in the way individual tax units report their adjusted gross income in response to tax law changes in the data analyzed by Piketty and Saez (2003).

[Place Figure 3 about here]

CREATING MORE CONSISTENT MEASURES OF TRENDS IN CROSS-GROUP INEQUALITY

While more sophisticated topcode correction methods can improve the accuracy of size-adjusted household income inequality calculations, as shown above, this is just one area where these newly available data and methods can improve calculations using public-use CPS data. Burkhauser and Larrimore (2008 and Forthcoming) demonstrate that not correcting for topcoding in the public-use CPS can also distort the size of the earnings gap across subsets of the population. This includes the gap in mean labor earnings between individuals by sex, race, education level, and disability statuses. Researchers using the public-use CPS data without our cell mean series will understate the mean labor earnings of each of these population groups. However, because there are more males than females with topcoded earnings, even consistent topcoding will suppress a larger percentage of male earnings, thus understating the male-female earnings gap. Similarly, White, highly educated, and nondisabled individuals are topcoded at higher rates than Black, less-educated, and disabled individuals, respectively, resulting in an understatement of the race, education, and disability earnings gaps by those using the public-use CPS.

As was the case in our income inequality comparisons above, using cell means largely corrects for these topcoding-based problems in the public-use CPS. When we use the public-use CPS together with our cell means, the earnings gaps we find within sex, race, education, and disability groups are nearly identical to those we find using the internal CPS.

INCORPORATING THE VARIANCES OF TOPCODED INCOMES

Using cell means with the public-use CPS allows researchers to more closely replicate results using the internal CPS. Moreover, researchers interested in more closely depicting the upper tail of the income distribution using the internal CPS can now do so by using the variances of topcoded incomes along with cell means. Burkhauser, Feng, and Larrimore (2008), using variance information from the internal CPS, move beyond imputing the same value for all topcoded individuals. They use a multiple imputation approach to create a distribution for the public-use CPS with the same mean and variance of topcoded incomes found in the internal CPS.

Figure 4, taken from Burkhauser, Feng, and Larrimore (2008), compares the log-variance of the income distribution using the public-use CPS without cell means (Public-NoMean) with cell means (Public-Mean) and with cell means and variances of topcoded incomes (Public-Variance) to the log-variance of incomes using the internal CPS data (Internal-Unadjusted). Since the Public-NoMean compresses the distribution by assigning all topcoded individuals' income equal to the topcode threshold, it is not surprising that this series understates the log-variance seen in the internal data. Using the Public-Mean series results in log-variance of income increases since it more accurately captures the level of topcoded income. However, it still understates the log-variance seen in the Internal-Unadjusted series. This is the case since it incorrectly assumes that all topcoded individuals have the same level of income, thus reducing the income variance. It is only when we impute topcoded incomes using both the cell mean and variance of topcoded incomes in the Public-Variance series that the log-variance of income from the public-use CPS closely matches the Internal-Unadjusted series.

[Place Figure 4 about here]

CONCLUSION

The CPS is the primary data source used by public policy researchers and administrators to investigate trends in income and its distribution. However, failure to control for inconsistent topcoding in the public-use CPS will lead researchers to understate the levels of U.S. average income and income inequality as well as distort their trends over time.

The series of papers discussed here, based on our ability to access internal CPS data, estimate and distribute the mean and variance of topcoded income values for each income source in these data.

Researchers using our cell mean and variance values together with the public-use CPS data can now closely approximate income and inequality levels and their trends based on the internal CPS. It is still important to consider internal censoring when using these data, especially when observing trends across 1992–1993. However, without making out-of-sample predictions about incomes censored in the internal data, these are the best estimates available. Moreover, they will be nearly as accurate as the official Census Bureau statistics that are published each year based on the internal CPS (U.S. Census Bureau, various years).

REFERENCES

Burkhauser, Richard V and Jeff Larrimore. 2008. “Using Internal Current Population Survey Data to Reevaluate Trends in Labor Earnings Gaps by Gender, Race, and Education Level.” Center for Economic Studies Working Paper CES-WP-08-18.

Burkhauser, Richard V and Jeff Larrimore. Forthcoming. “Trends in the Relative Household Income of Working-Age Men With Work Limitations: Correcting the Record Using Internal Current Population Survey Data.” *Journal of Disability Policy Studies*

Burkhauser, Richard V, Shuaizhang Feng, and Jeff Larrimore. 2008. “Measuring Labor Earnings Inequality Using Public-Use March Current Population Survey Data: The Value of Including Variance and Cell Means When Imputing Topcoded Values.” Center for Economic Studies Working Paper CES-WP-08-38.

Burkhauser, Richard V, Shuaizhang Feng, and Stephen Jenkins. 2009. ”Using the P90/P10 Ratio to Measure U.S. Inequality Trends With Current Population Survey Data: A View From Inside the Census Bureau Vaults.” *The Review of Income and Wealth* 55(1): 166–185.

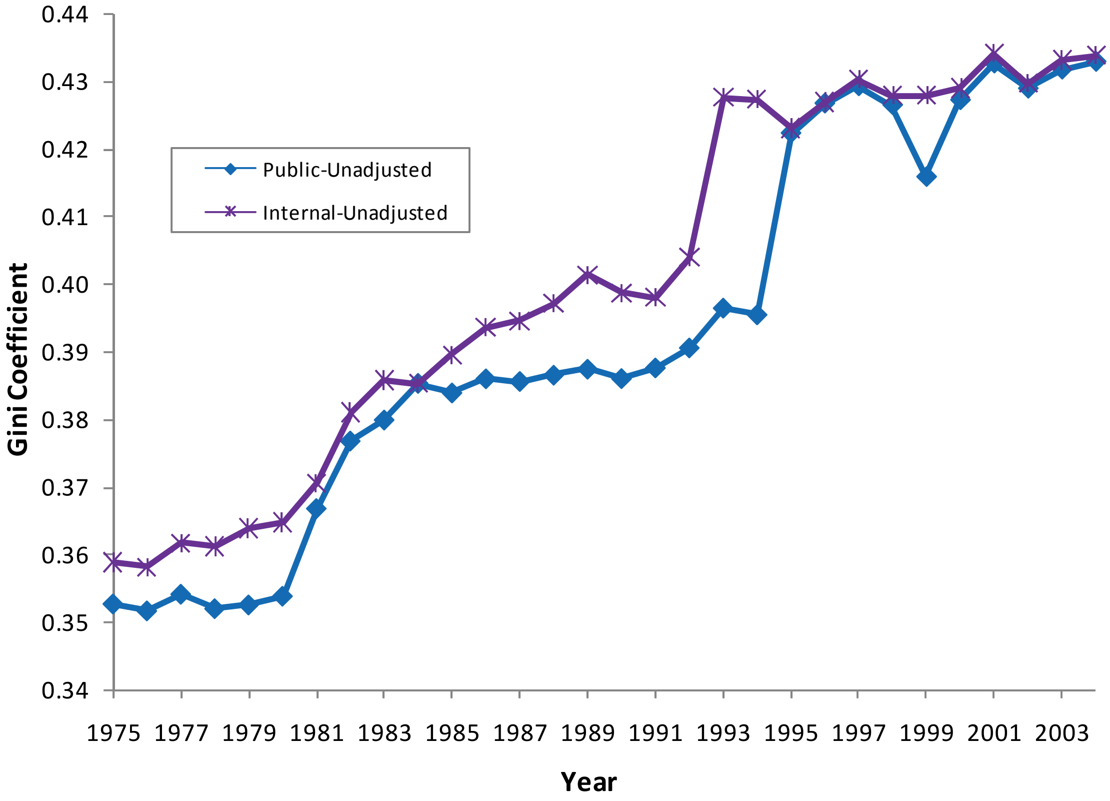
Burkhauser, Richard V, Shuaizhang Feng, Stephen Jenkins, and Jeff Larrimore. 2008. ”Estimating Trends in United States Income Inequality Using the March Current Population Survey: The Importance of Controlling for Censoring." Center for Economic Studies Working Paper CES-WP-08-25.

Larrimore, Jeff; Richard V Burkhauser; Shuaizhang Feng, and Laura Zayatz. 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.

Piketty, Thomas and Emmanuel Saez. 2003. ”Income Inequality in the United States, 1913–1998.” *Quarterly Journal of Economics*, 118(1): 1–39.

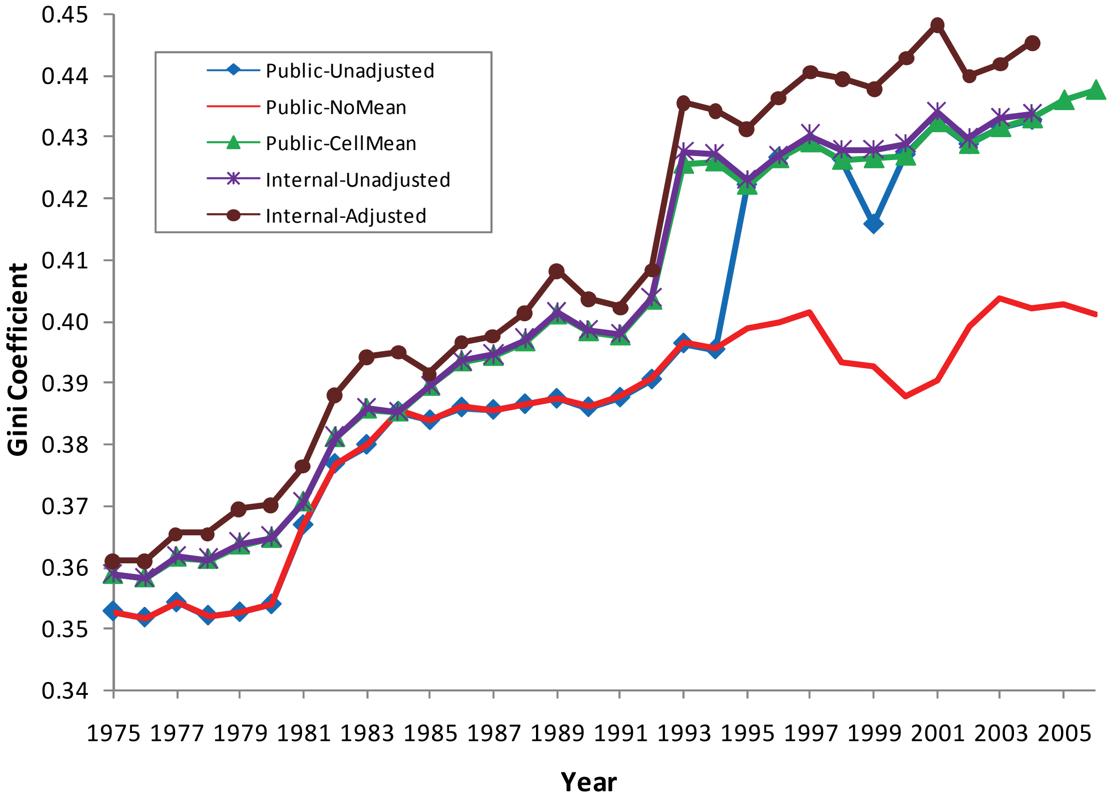
U.S. Census Bureau. Various years. *Income, Poverty, and Health Insurance Coverage in the United States*. (August), Current Population Reports, P60 series, Washington, DC: U.S. Government Printing Office.

Figure 1: Income Inequality Is Understated When Using Unadjusted Public-Use CPS Data: 1975 to 2004



Source: Burkhauser, Feng, Jenkins, and Larrimore (2008).

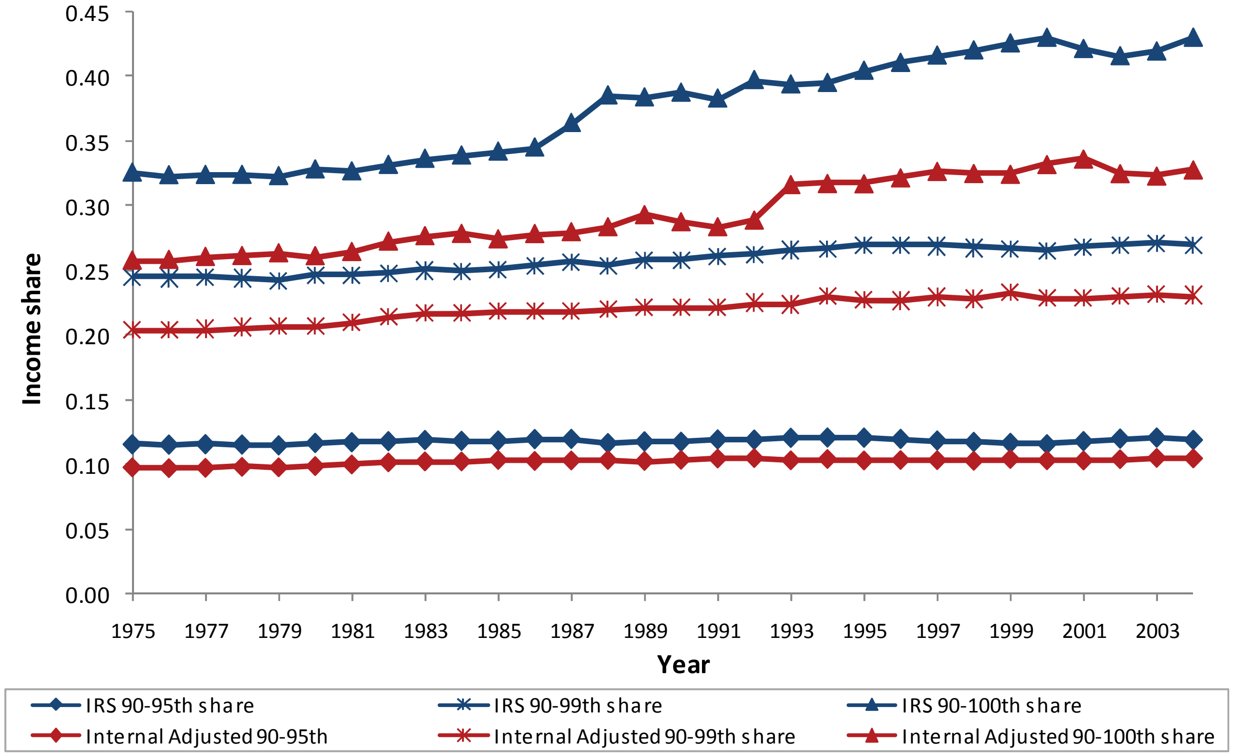
Figure 2: Measured Income Inequality Increases When Using Alternative Topcode Correction Methods Compared to the Unadjusted Public-Use CPS Data: 1975 to 2006



Note: Internal data were not available for years after 2005.

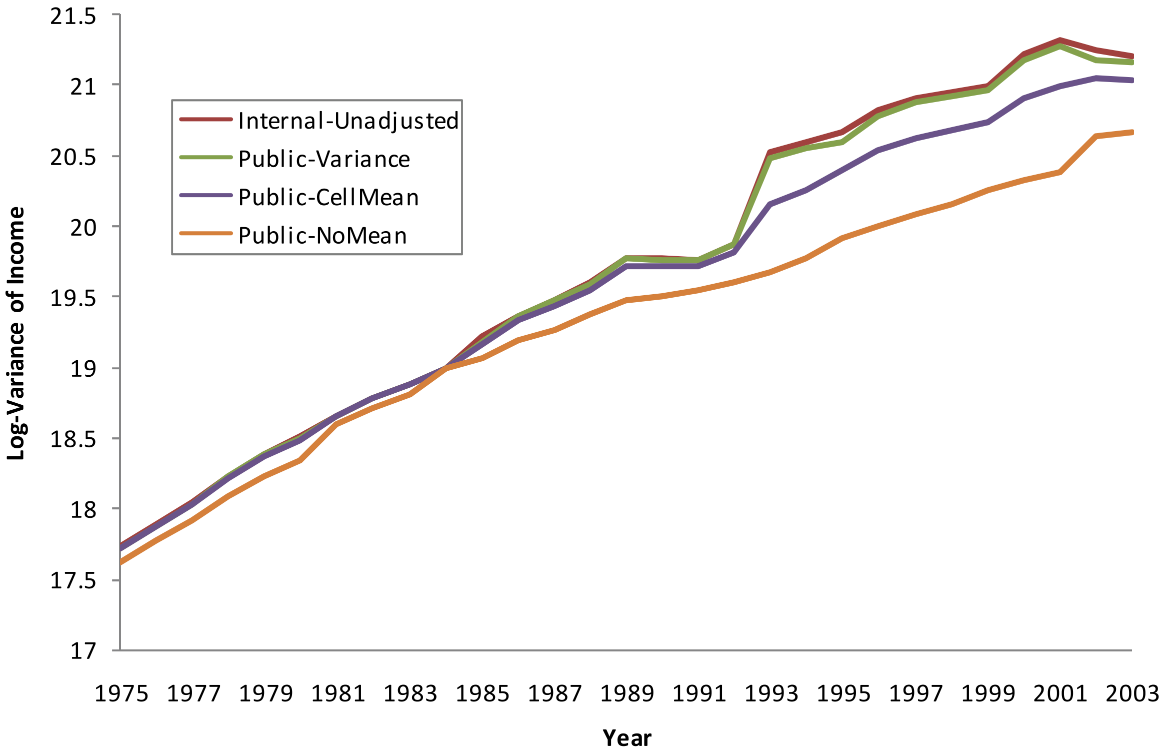
Source: Burkhauser, Feng, Jenkins, and Larrimore (2008).

Figure 3. Similar Shares of Income Held by the Top Part of the Distribution—IRS vs. CPS: 1975 to 2004



Sources: Burkhauser, Feng, Jenkins, and Larrimore (2008) and <http://elsa.berkeley.edu/~saez/>.

Figure 4: Imputed Cell Means and Variances Better Capture the Log-Variance of Income: 1975 to 2004



Source: Burkhauser, Feng, and Larrimore (2008).

Table 1: Definitions for Income Distribution Series by Source and Censoring Method

Acronym	Source	Method for Addressing Censoring Issues
Internal-Unadjusted	Internal	Uses internal data as provided in Census Bureau files, without any adjustments
Internal-Adjusted	Internal	Topcoded observations replaced by imputations derived from multiple imputation model fitted to internal data; inequality estimates derived using multiple imputation combination methods.
Public-Unadjusted	Public-Use	Uses public-use data as provided in Census Bureau files; includes Census Bureau cell mean imputations for topcoded observations from 1995 onwards
Public-CellMean	Public-Use	Uses public-use data as provided in Census Bureau files; includes cell mean imputations for topcoded observations for all years
Public-NoMean	Public-Use	Uses public-use data as provided in Census Bureau files, except that no cell mean imputations used for any year (topcoded values used 'as is').
Public-Variance	Public-Use	Uses public-use data as provided in Census Bureau files; topcoded observations are replaced by imputations derived from a Stoppa imputation model fitted to the mean and variance of topcoded incomes from the internal data

Sources: Burkhauser, Feng, Jenkins and Larrimore (2008) and Burkhauser, Feng and Larrimore (2008).