Burkhauser, Richard V.; Feng, Shuaizhang; Jenkins, Stephen P.; Larrimore, Jeff

Working Paper

Estimating trends in US income inequality using the Current Population Survey: the importance of controlling for censoring

ISER Working Paper Series, No. 2008-25

Provided in Cooperation with:
Institute for Social and Economic Research (ISER), University of Essex

Suggested Citation: Burkhauser, Richard V.; Feng, Shuaizhang; Jenkins, Stephen P.; Larrimore, Jeff (2008) : Estimating trends in US income inequality using the Current Population Survey: the importance of controlling for censoring, ISER Working Paper Series, No. 2008-25, University of Essex, Institute for Social and Economic Research (ISER), Colchester

This Version is available at:
https://hdl.handle.net/10419/92206

ISER Working Paper Series

Estimating trends in US income inequality using the Current Population Survey: the importance of controlling for censoring

Richard V. Burkhauser
Cornell University

Shuaizhang Feng
Shanghai University of
Finance and Economics

Stephen P. Jenkins
ISER, University of Essex

Jeff Larrimore
Cornell University

INSTITUTE FOR SOCIAL
& ECONOMIC RESEARCH

No. 2008-25
September 2008

E·S·R·C
ECONOMIC
& SOCIAL
RESEARCH
COUNCIL

www.iser.essex.ac.uk

Non-technical summary

There is intense public interest in the spread of incomes in society between the rich and the poor and those in the middle, and whether income gaps have been growing larger over time. Although it is widely agreed that income inequality in the USA rose during the 1980s, there is less agreement about what happened during the 1990s and later. There are two main sources of data that researchers might use to investigate this topic. The first is the income information collected from respondents to large-scale surveys of households, and the second is administrative data on income that are kept by the US income tax authorities. Both sources have some advantages and some disadvantages, and so complement each other.

We analyze trends in US income inequality between 1975 and 2004, using data from the March Current Population Survey (CPS), the source most commonly used to address this topic. An important limitation of the CPS data that are made available to researchers (the 'public use' data) is that they are censored ('topcoded'). In order to protect respondent confidentiality and for other reasons, the very richest incomes reported in interviews are not included in the data; instead they are replaced in the public use data with some lower value (the 'topcode'). Not only does censoring reduce measured inequality, but changes over time in topcoding practices complicate assessments of trends in inequality. To investigate the effects of censoring problems in public-use CPS data, we secured unprecedented access to CPS internal data. Although there is some censoring in the internal data as well, its prevalence is relatively small and, moreover, we employ a statistical technique that enables us to impute plausible values to the small number of the very richest people for whom actual incomes are not included in the internal data. So, with access to the public use and internal CPS data, we can examine how trends in income equality changed over time, and check the robustness of our findings using a number of different data series. We also use several methods for summarizing inequality itself.

We find, as others have, that income inequality increased during the 1980s. However, we also show that the upward trend in income inequality slowed significantly after 1993.

To investigate the robustness of the slowdown finding, we compared our CPS-derived results with those derived from information about incomes derived from administrative record data from the US Internal Revenue Service and reported in research by Thomas Piketty and Emanuel Saez. The focus is on trends in top income shares – the share of total income that is held by the richest 1%, 2%, 5%, and so on. We show that our CPS estimates of trends in top income shares match the estimates of trends reported by Piketty and Saez, except for within the top 1% of the distribution. Thus, we argue that, if income inequality in the USA has increased substantially since 1993, such increases are confined to this very highest income group.

Estimating Trends in US Income Inequality Using the Current Population Survey: The Importance of Controlling for Censoring

Richard V. Burkhauser
(Cornell University)

Shuaizhang Feng
(Shanghai University of Finance and Economics)

Stephen P. Jenkins
(ISER, University of Essex, email: stephenj@essex.ac.uk)

Jeff Larrimore
(Cornell University)

August 20, 2008

Abstract

Using internal and public use March Current Population Survey data, we analyze trends in US income inequality (1975–2004). Using a multiple imputation approach where values for censored observations are imputed using draws from a Generalized Beta distribution of the Second Kind, we find that the upward trend in income inequality significantly slowed after 1993. Our results closely match the income share trends reported by Piketty and Saez (2003) except for within the top 1 percent of the distribution. Thus, we argue that if inequality has increased substantially since 1993, such increases are confined to this very high income group.

Key words: US income inequality, time trend, CPS, censoring, topcoding, inequality
JEL classifications: D31, C81

Acknowledgements

The research in this paper was conducted while Burkhauser, Feng and Larrimore were Special Sworn Status researchers of the U.S. Census Bureau at the New York Census Research Data Center at Cornell University. Conclusions expressed are those of the authors and do not necessarily reflect the views of the U.S. Census Bureau. This paper has been screened to ensure that no confidential data are disclosed. Supports 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. Jenkins's research was supported by core funding from the University of Essex and the UK Economic and Social Research Council for the Research Centre on Micro-Social Change and the United Kingdom Longitudinal Studies Centre. We thank Lisa Marie Dragoset, Arnie Reznek, Laura Zayatz, the Cornell Census RDC Administrators, and all their U.S. Census Bureau colleagues who have helped with this project. This paper has also benefited from comments made by Peter Gottschalk, Bruce Meyer, and other participants at the Empirical Economics Workshop of Shanghai University of Finance and Economics, Shanghai, May 2008. The paper has also been released as NBER Working Paper No. 14247.

Introduction

The public use version of the March Current Population Survey (CPS) is the primary data source used by public policy researchers and administrators to investigate trends in average income and its distribution in the United States as well as in cross-national comparisons of income inequality with other countries. (For systematic reviews of the cross-national income inequality literature, see Atkinson, Rainwater and Smeeding, 1995, Gottschalk and Smeeding, 1997, Atkinson and Brandolini, 2001. For more recent examples of the use of the public use CPS in measuring income inequality trends in the United States see: Burkhauser, Couch, Houtenville and Rovba 2003-2004, Gottschalk and Danziger, 2005 and Kruger and Perri, 2006.) The consensus of this research, based on public use CPS data is that income inequality in the United States increased substantially in the 1970s and 1980s and also increased relative to other OECD countries.

Despite the widely held view that income inequality has increased substantially since the 1980s, most of the evidence of a large increase in income inequality since 1993 has come from Internal Revenue Service (IRS) administrative record data based on personal adjusted gross income that the IRS has made available to the research community in tabular form (Piketty and Saez, 2003). In contrast, we argue that yearly income inequality increases are small, and significantly smaller than the increases over the nearly two decades of CPS data available to us prior to 1993.

We compare our CPS-based estimates of inequality trends with those of Piketty and Saez (2003) based on IRS data. When, like Piketty and Saez, we focus on the share of reported income held by subgroups within the richest tenth of the distribution, our trends fairly closely match their trends, except for the top 1 percent of the distribution over the period 1975–2004. On the basis of these findings, we argue that if income inequality has been substantially increasing since 1993 in the USA, such increases have been confined to the very upper tail of the income distribution.

Our analysis derives from unprecedented access to internal CPS data. To protect the confidentiality of its respondents, the Census Bureau censors each source of income of individuals above specified topcoded levels in the public use data made available to the research community. However, for its official work, researchers at the Census Bureau have access to the internal March CPS that is less severely censored. For instance, the yearly income and income inequality values reported in U.S. Census Bureau (various years) are based on the internal March CPS data. We use these data too.

We analyze trends in the inequality of size-adjusted pre-tax post-transfer household income in the USA between 1975 and 2004. Estimates of inequality levels and trends derived from CPS internal data are compared with estimates from several series derived from public use data. In particular, using the internal data, we derive a cell mean series in which topcoded incomes in the public use data, for each year back to 1975, are replaced by the mean of all income values above the topcode for any income source of any individual in the public use that has been topcoded.[1] This cell mean series can be used in conjunction with cell means provided by the Census Bureau for later years to create a complete set of cell means for topcoded observations. See Larrimore, Burkhauser, Feng, and Zayatz (forthcoming) for a detailed discussion of the creation of this cell mean series and all cell mean values.

When we use the public use data augmented with imputations for topcoded observations from the extended cell mean series, we closely match yearly mean income and income inequality levels and trends in the United States population derived from internal data.[2] Internal data are themselves censored albeit to a substantially smaller extent than the public use data, and there are time-inconsistencies in censoring practices in the internal data. However, we show using consistent censoring methods that the trends in inequality derived from our public use extended cell mean series are not significantly affected by time-inconsistent censoring levels. The exception is between 1992 and 1993, when the public use extended cell mean series overstates the rise in inequality because it does not take full account of the significant increases in internal censoring levels that took effect in 1993, and other CPS redesign aspects.

Because consistently censored public use and internal data do not contain information about the very highest incomes, they understate the level of income inequality. However, once the break between 1992 and 1993 is controlled for, income inequality based on all these

[1] Each CPS survey measures income from the previous calendar year. In this paper, all references are to the income year, so when we discuss the year 1975, this refers to income from various sources that members of the household received in 1975 reported at the March 1976 Current Population Survey interview.

[2] In 2006, we were granted permission to use the internal March CPS to test the sensitivity of measured income inequality in the public use CPS. We were given access to internal March CPS records from 1975–2004. These data include information on income above the public use topcode thresholds for respondents in the March CPS survey up to the *processing limit* in the March CPS data. The internal March CPS data the Census Bureau uses to produce the statistics in its official Census publications are subject to these same processing limits. However these processing limits are lower than the *data collection limits* for income in the March CPS, which are the limits on the actual values collected in the March CPS. (See Welniak, 2003 for a more complete discussion of the processing limits and data collection limits.) Although we had access to the same income information the Census Bureau uses to produce its official Census publications, since our research project was limited to how topcoding in the public files and censoring in the internal files affect measurement of inequality in the USA, the files do not contain the responses to all of the non-income related questions in the March CPS dataset. These omissions occur because the U.S. Census Bureau limits internal data availability to data within the scope of the project.

data series yield the same trends, namely a rise in inequality between 1975 and 1992 followed by a significantly smaller increase in inequality after 1993. Our key findings are generally robust to whether income inequality is summarized using the Gini coefficient or three other widely-used indices.

Because of the censoring in the internal CPS data, even our extended cell mean series does not incorporate information about the very highest incomes. To address this issue, we use a multiple imputation approach in which, for each year, values for censored observations in the internal data are imputed using draws from a Generalized Beta distribution of the Second Kind (GB2) fitted to internal data. Using the augmented internal data, we investigate the extent to which the systematic exclusion of the top part of the distribution affects estimates of the level and trends in inequality. Unsurprisingly, we find that compared to estimates derived from the multiple imputation approach, the unadjusted internal data as well as the consistently censored public use and internal data all understate the level of inequality in all years. However, all the series reveal 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.

Finally, we compare our CPS-based estimates of trends in top income shares based on distributions of size-adjusted pre-tax post-transfer household income distributions to the estimates of top income shares derived by Piketty and Saez (2003) from IRS administrative data on adjusted personal gross tax unit income. Both our and their estimates of the income share of the 90^{th}–95^{th} percentile group have a relatively flat trend during the period 1975–2004, although the Piketty and Saez (2003) values are slightly higher in level. Similarly, for the shares of the 95^{th}–99^{th} percentile group, despite slight differences in levels, the two series exhibit remarkably similar trends over the 30 year period. In contrast, while the share of income held by the richest 1 percent increased substantially over this period according to both our CPS-based and the IRS-based Piketty and Saez (2003) estimates, their estimates of the size of the income shares and of the magnitude of the increase over time are much larger than ours.

In the next section, we explain the nature of censoring in the CPS public use and internal data, and explain the several methods by which we address this issue. In particular, we explain how we derive a public use data series augmented by cell mean imputations for topcoded observations based on the censored internal data. We also explain how we derive other data series in which public use and internal data are augmented by imputations for censored observations using draws from a parametric income distribution model fitted to the

data. And we explain our method of 'consistent' censoring. At the heart of the paper are comparisons of estimates derived from the various CPS data series of income inequality levels and trends for the period 1975–2004. The fourth section provides a comparison of CPS-based estimate of trends in top income shares with the trends reported by Piketty and Saez (2003) using IRS administrative record data. The final section of the paper presents a summary and conclusions.

Throughout the paper, income is defined as pre-tax post-transfer household income excluding capital gains, adjusted for differences in household size using the square root of household size.[3] Each individual is attributed with the size-adjusted income of the household to which he or she belongs. Income refers to income for the calendar year preceding the March interview. We convert the small number of negative and zero household income values each year to one dollar prior to our calculations because a number of inequality indices are defined only for positive income values. Our samples for each year are all individuals in CPS respondent households, excluding individuals in group quarters or in households containing a member of the military. All statistics are calculated using the relevant CPS sampling weights.

Censoring in the March CPS

Topcoding in CPS public use data

In the March CPS, a respondent in each household is asked a series of questions on the sources of income for the household. Starting in 1975, respondents reported income from 11 sources and since 1987 they have done so for income from 24 sources. See Appendix Table 1 for a list of these income sources. Rather than simply topcoding high total household income values in the public use data, the US Census Bureau topcodes high values for each source of household income. See Larrimore, Burkhauser, Feng, and Zayatz (forthcoming) for a full list of topcode values over time in the public use data. Prior to 1995, the Census Bureau assigned the topcode value from that source of income to all topcoded income values in the public use data. Since then, they have substituted a cell mean value derived from the internal data to each topcoded value in the public use data.

[3] We follow the same procedure for generating size-adjusted household income as discussed in Burkhauser and Larrimore, 2008. These procedures, particularly the use of the square root of household members to adjust for household size are standard in the international comparative income distribution literature: see for instance Atkinson, Rainwater and Smeeding (1995). Its use is increasing in studies of US income distribution trends. See Burkhauser, Couch, Houtenville and Rovba (2004), Gottschalk and Danziger (2005), and Burkhauser, Osaki, and Rovba (2008).

Because the Census Bureau cell means series starts in 1995, using the public use data without correcting for this major change in the values included for the highest incomes, source by source, results in a significant increase in measured income and income inequality in 1995 and subsequently simply because of the closer correspondence between the true reported income and the value included in the public use file. Hence, whereas the use of cell mean imputations after 1995 makes the public use data better conform to the internal data, not taking this improvement in measurement into account overestimates how much actual income increased in 1995 among those at the highest income levels and the overall levels of income inequality. An additional complication is that household income is the aggregation of multiple income sources (income types and household members), each of which may be topcoded. As a result, the prevalence of topcoding in household income is significantly greater than for any particular income source, and topcoded household income values are not necessarily the highest incomes – they may occur throughout the income distribution. Thus using the ratio of the 90th percentile to the 10th percentile (the 'P90/P10 ratio') to minimize the impact of topcoding on inequality estimates will not be entirely successful: see Burkhauser, Feng, and Jenkins (2007).

The major change in the public use data in 1995 is a specific example of the more general problem that income topcoding presents researchers interested in measuring levels and trends in the income and income distribution of the US population. Inconsistently-defined topcode values lead to artificial increases or decreases in mean income and income inequality, since different fractions of the population are subject to topcoding each year.[4] This is a legitimate concern since topcodes in the public use data have, in general, increased in a non-systematic manner over its history, and part of the apparent trend in average income and income dispersion over time in the uncorrected data is caused by topcoding capturing a larger portion of the income distribution. Using cell mean imputations for topcoded observations substantially alleviates this problem after 1995, since cell means provide the information necessary to capture the income distribution represented by the internal data. As a result, even if public topcodes are inconsistently adjusted in the public use data, mean income for the entire population in the internal data can be matched with the public use data using cell mean imputations.

Despite the Census Bureau's attempt to alleviate the problem of topcoding, their cell means have generally been ignored by researchers studying long-term income trends in the

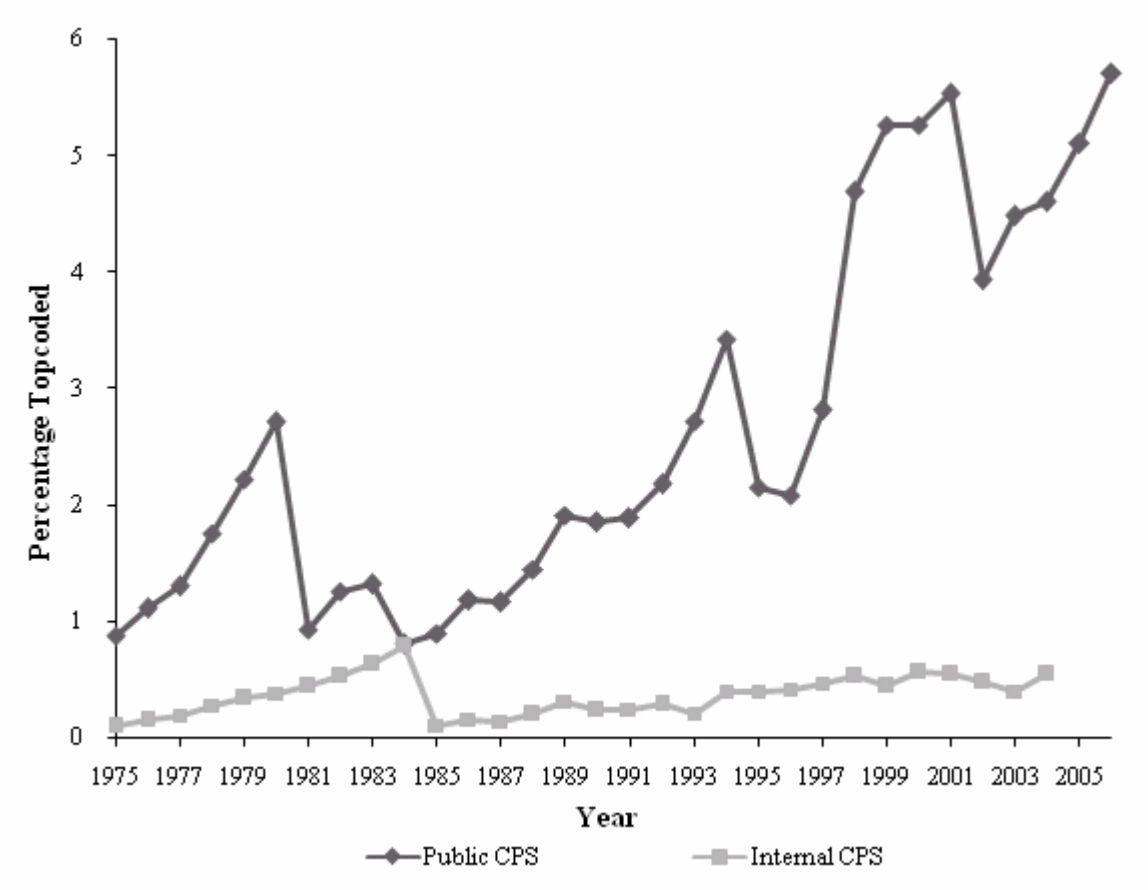
[4] See Levy and Murnane (1992) for an early review of the income distribution literature and a more formal statement of this problem.

USA, since to do otherwise exacerbates time-inconsistencies that arise from using unadjusted public use data for the pre-1995 period and CPS data with cell mean imputations thereafter. Instead, researchers analyzing CPS data that includes years prior to 1995 use other methods of controlling for inconsistent topcoding in the public use data. These methods include measuring inequality with P90/P10 ratios (e.g. Danziger and Gottschalk, 1993; Gottschalk and Danziger, 2005; Daly and Valletta, 2006); artificially truncating the data by removing the highest and lowest two percent of observations (e.g. Welch, 1999); or artificially lowering the topcodes to create a series in which for each income source, a constant percentage of that source's income distribution is censored in the data for each year (e.g. Karoly and Burtless, 1995; Burkhauser, Butler, Feng, and Houtenville, 2004). This is referred to as the consistent censoring method. While each of these methods is preferable to using unadjusted data, each has drawbacks.

Burkhauser, Feng, and Jenkins (2007) show that using P90/P10 ratios does not completely alleviate topcoding problems, since incomes are topcoded by income source (not by total income) and so there are topcoded incomes below the 90^{th} percentile. Additionally, the P90/P10 ratio captures different aspects of the income distribution than widely-used measures of dispersion such as the Gini coefficient or members of the Generalized Entropy family.

Consistent censoring of the public use data can also cause problems. The percentage of individuals with topcoded incomes, whether based on their own income or the income of others in their household each year in the public use data, has grown sporadically over time from just under 1 percent in 1975 to nearly 6 percent in 2006: see Figure 1. Unless public use data topcodes are systematically increased over time as income increases, these topcodes cut deeper into the income distribution. In that case, consistent censoring methods do an increasingly poorer job of capturing trends at the top of the income distribution as they remove larger fractions of the population to maintain their consistency.

Figure 1
Percentage of Individuals with Censored Household Income in March CPS, by year



Internal data were not available for years after 2004.
Source: Authors' calculations from internal and public use data files of March CPS.

Censoring in CPS internal data

Using public use data with cell mean imputations cannot solve the problem of censoring entirely, because the internal data from which the cell means are derived are themselves censored, albeit to a lesser degree than the public use data: see Figure 1.

Censoring in the internal data was initially implemented due to data-storage limitations in the computing systems of the 1970s which necessitated truncating written records to 5 digits. This was most binding on wage income, self-employment income and farm income. While these data storage limitations are no longer a constraint, the Census Bureau continues its internal censoring practice, partly due to concerns about data reliability of individuals who report an extremely high income value and partly due to concerns about confidentiality.

In 1985 censoring limits were raised to $250,000 on each source of labor earnings and internal censoring fell close to zero. Subsequent increases in censoring values have kept the percentage of censored individuals in the internal data well below 1 percent although increases in censoring from non-labor income sources in the past few years has increased the share of censored individuals somewhat since 1994. See Larrimore, Burkhauser, Feng, and Zayatz (forthcoming) for the censoring levels over time for each income source in the internal data and see Ryscavage (1995), Jones and Weinberg (2000), Welniak (2003), and Burkhauser et al. (2004), for earlier discussions of the problems of censoring in the public use and internal data.

Because we had access to the internal CPS data, we are able for the first time to correct for censoring in these data back to 1975, and hence provide a more consistent measure of trends in the income distribution and compare these trends with those found with the public use data.

A Consistent Censoring Approach

Consistent censoring (e.g. Karoly and Burtless, 1995; Burkhauser, Butler, Feng, and Houtenville, 2004) creates censoring points for each income source that captures the same percentile of that source's income in every year. Because total household income is derived as the sum of all of the individual sources of income in a household, it is necessary to create a consistent censoring point for each of these sources of income over all years. We do so by finding the year at which the censoring point for a given source is at the lowest point in the distribution of income within that source and then using that percentile as our cutoff point for that source for all years. Since the sources of income reported by the Census Bureau changed in 1987, we recode income and censoring points for years after this time to the pre-1987 sources in order to create a time consistent series across the entire 30 year period between 1975 and 2004. Appendix Table 2 reports the percentage of the distribution affected by our censoring points and the year in which the most restrictive censoring points occurred for each source of income in both the public use and internal data sets.

The consistent censoring approach makes no parametric assumptions about the underlying income distribution, and maximizes the use of the common proportion of the available income information. Nevertheless, by construction, the approach ignores the very highest incomes in the distribution each year, and so it will underestimate levels of inequality. However, the fractions omitted do not vary over time, and so this approach provides time-

consistent trends of inequality for the part of the distribution below a certain fixed percentile that it does capture.

A Multiple Imputation Approach

To correct for censoring and also capture the whole distribution, we develop the imputation idea in a different direction. The key differences from the cell mean approach are, first, we attempt to account for censoring in the internal data (as well as in public use data) and, second, we derive imputations for censored values using a parametric model of the income distribution (since, by definition, the 'true' uncensored underlying income in the internal data are not available to us). Third, while the cell mean approach leads to a single imputation for each topcoded value, we derive multiple imputed values using a suitable randomization procedure which leads to multiple distributions for each year, and account for potential stochastic imputation error by averaging estimates (hence the label 'multiple imputation').

The parametric model of the income distribution used for the imputation model is the Generalized Beta of the Second Kind (GB2). The GB2 is a flexible form widely used in the income distribution literature, and a number of studies have shown that it fits income distributions extremely well across different times and countries: see e.g. Bordley, McDonald and Mantrala (1996), Brachmann, Stich and Trede (1996), and Bandourian, McDonald, and Turley (2002). Feng, Burkhauser, and Butler (2006) fitted GB2 distributions to labor earnings using public use CPS data and argue that their Gini coefficients calculated from the fitted GB2 distributions provide more plausible estimates of inequality trends than do the Gini coefficients derived from unadjusted internal data reported by the Census Bureau.

Our multiple imputation approach differs from earlier studies that used parametric models to derive imputed values for topcoded incomes: see Fichtenbaum and Shahidi (1988) and Bishop, Chiou, and Formby (1994). First, those authors fitted a one-parameter Pareto distribution to the upper ranges of the income distribution, rather than the more flexible four-parameter GB2 model. Second, our approach takes account of the fact that, rather than being common to all individuals, censoring levels vary across individuals (because censoring is done at the level of each income source rather than at the level of aggregate household income). Third, by using multiple imputation methods, rather than a single imputation, we account for the variability intrinsic in the imputation process.

Our multiple imputation approach involves five steps. First, for each year's data, we fit a GB2 distribution by maximum likelihood, accounting for individual level right-censoring. In fitting the GB2 distribution, we model individual household size-adjusted income at the

aggregate level, not for each income source separately. Doing so the other way would require modeling the joint distribution of 24 income sources (11 before 1987) for various members of each household and as a result our approach is more parsimonious and tractable. To ensure that model fit is maximized at the top of the distribution, the GB2 is fitted using observations in the richest 70 percent of the distribution only (using appropriate corrections for left truncation in the ML procedure). Second, for each observation with a censored income, we draw a value from the income distribution that is implied by the fitted GB2 distribution, using an appropriate stochastic procedure. Third, using the distribution comprising imputations for censored observations and observed incomes for non-censored observations, we estimate our various inequality indices. Fourth, we repeated steps 2 and 3 one hundred times, and finally, we combine the one hundred sets of estimates from each of the one hundred data sets for each year using the 'averaging' rules proposed by Rubin (1987) and modified by Reiter (2003) for the case of partially synthetic data. See Jenkins, Burkhauser, Feng and Larrimore (2008) for further details of our multiple imputation approach.

In what follows we estimate inequality levels and trends using different series defined according to the method by which we take account of censoring, and whether the basis source is the public use or internal data. The series, and the acronyms by which we refer to them subsequently, are summarized in Table 1.

Table 1
Income Distribution Series Derived from the March CPS, by Source and Method for Addressing Censoring

Acronym	Source	Method for Addressing Censoring Issues
Internal-CC	Internal	Consistently Censoring
Internal-Unadjusted	Internal	Uses internal data as provided in Census Bureau files, without any adjustments
Internal-MI	Internal	Topcoded observations replaced by imputations derived from GB2 imputation model fitted to internal data; inequality estimates derived using multiple imputation combination methods.
Public-CC	Public Use	Consistently Censoring
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-CM	Public Use	Uses public use data as provided in Census Bureau files; includes cell mean imputations for topcoded observations for all years
Public-NoCM	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-MI	Public Use	Topcoded observations replaced by imputations derived from GB2 imputation model fitted to public use data; inequality estimates derived using multiple imputation combination methods.

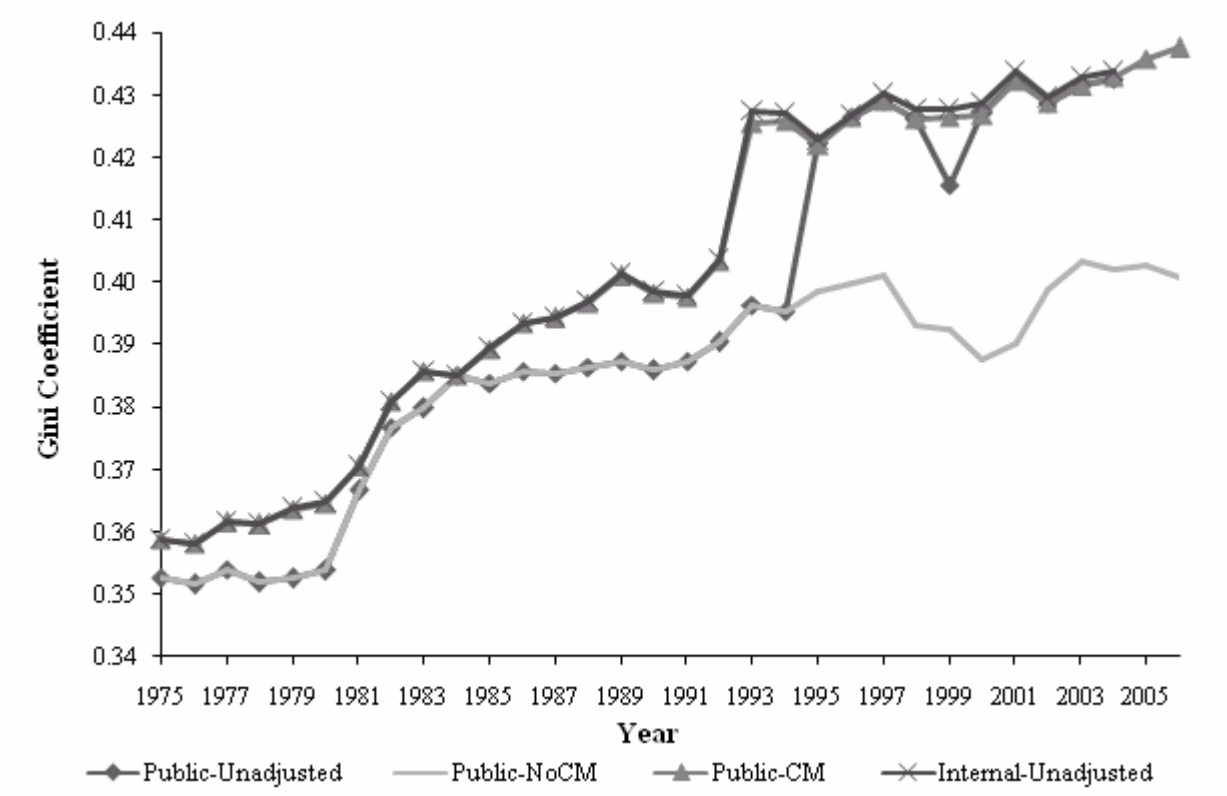
See main text for further details of the methods used to address censoring issues.

Estimates of inequality levels and trends, 1975–2006

The revised cell mean series (Public-CM)

Figure 2 reports our estimates of income inequality levels and trends derived from four series: Public-NoCM, Public-Unadjusted, Public-CM, and Internal-Unadjusted. Inequality is summarized using the Gini coefficient. Because public use data are topcoded, mean household income and Gini coefficients derived from them will be understated relative to those estimated from the internal data unless cell mean imputations are used, since cell means provide more information about individuals at the top of the income distribution.

Figure 2
Gini Coefficient Estimates for Four Censoring Adjustment Methods, 1975–2006



Internal data were not available for years after 2004.
Source: Authors' calculations from internal and public use data files of March CPS.

In all years, the Gini estimates based on the Public-NoCM series are below those from the Internal-Unadjusted series. To correct for this difference (and for several other reasons), the Census Bureau revised its topcoding procedures in income year 1995. The result is reflected in the difference between the Gini estimates derived from the Public-NoCM and Public-Unadjusted series. Prior to 1995 the series are identical. After 1995 the Public-Unadjusted Gini values equal the Internal-Unadjusted Gini values, and are dramatically greater than those derived from the Public-NoCM series. (Note that this is not the case in 1999. We believe this is the result of an uncorrected error in the Census Bureau cell mean series.) While the Gini values from the Public-Unadjusted series now better represents the Gini values derived from the Internal-Unadjusted data, anyone uncritically using Gini values from the Public-Unadjusted data to describe changes in income inequality in the United States would greatly exaggerate its increase in 1995.

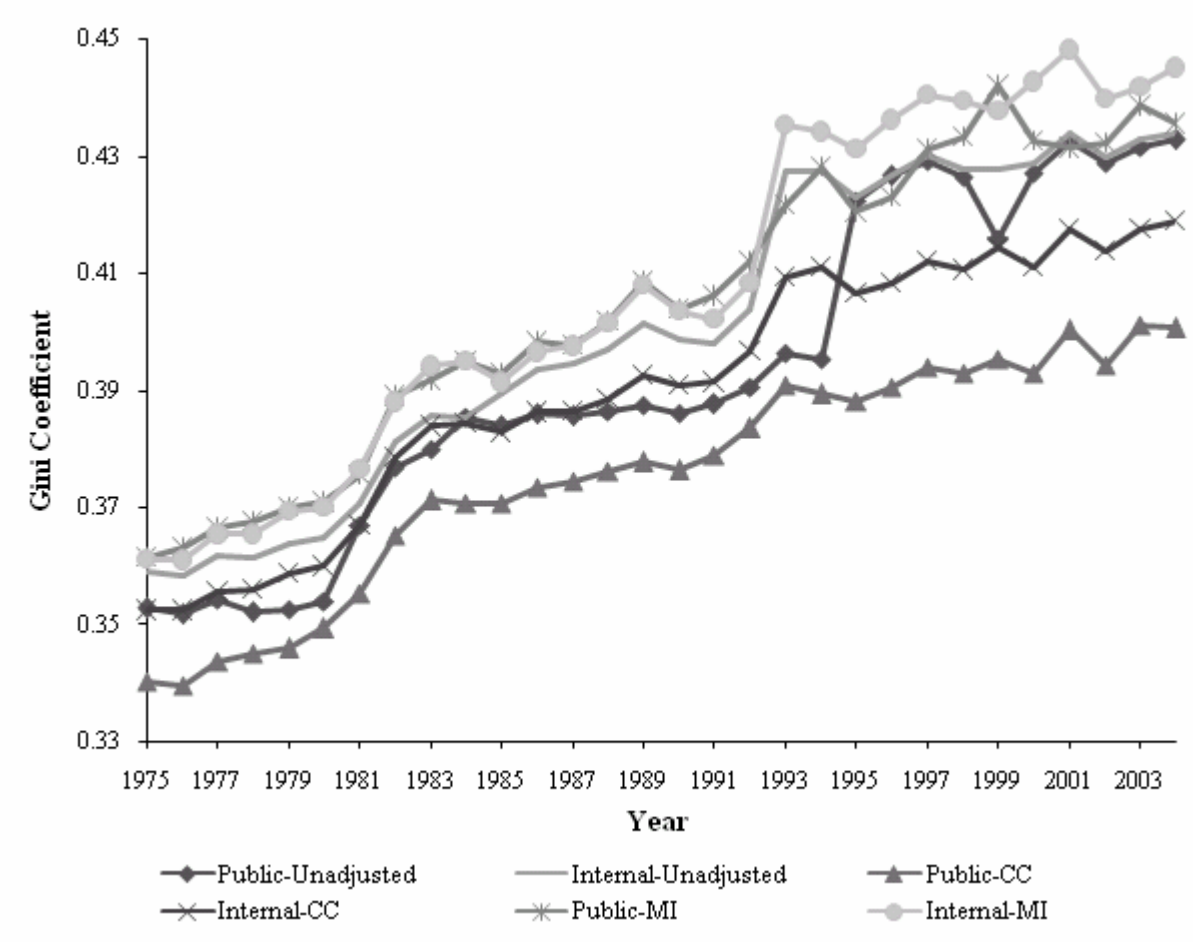
Our Public-CM series corrects this problem – the Gini estimates derived from these data match those derived from the Internal-Unadjusted data in all years including 1999: see Figure 2. Thus the addition of our cell mean imputations to the CPS public use data allows researchers without access to the Internal-Unadjusted data to produce the same inequality levels and trends found in the Internal-Unadjusted data. But anyone uncritically using Gini values based on the Internal-Unadjusted data to describe changes in income inequality would greatly exaggerate its increase in 1993 when major changes occurred in the way the CPS was collected including significant changes in the censoring points in the internal CPS data**.**

Long term trends in US income inequality

We now extend our analysis of long term trends using additional data series. There are six in total: Public-Unadjusted; Public-CC; Public-MI, Internal-Unadjusted; Internal-CC, and Internal-MI. Figure 3 depicts the trends over the period 1975–2004 in the Gini coefficients derived from each of the series. We stop at 2004 because that is the last year of internal CPS data to which we had access.

As discussed above, the use of cell means since 1995 in the Public-Unadjusted data allows researchers to estimate Gini coefficients that closely match those estimated using the Internal-Unadjusted data since then. But using the Public-Unadjusted Gini estimates to describe trends in income inequality before and after 1995 distorts these trends. When we consistently censor public use data (Public-CC series), the Gini estimate for each year is below the corresponding estimate from the Public-Unadjusted data: see Figure 3. But the artificial jump in 1995 in the Public-Unadjusted Gini values is gone and the trend in Public-CC Gini values more closely follows the trends estimated using the Internal-Unadjusted data, including a noticeable rise in inequality between 1992 and 1993.

Figure 3
Gini Coefficient Estimates for Six Censoring Adjustment Methods, 1975–2004



Source: Authors' calculations from internal and public use data files of March CPS.

When we repeat this exercise by consistently censoring internal data (Internal-CC series), once again the Gini estimates are smaller than those based on the Internal-Unadjusted data but they are much closer to those estimated from the Internal-Unadjusted data than those based on the Public-CC data. This is unsurprising, since the Internal-CC data now reach a higher percentile of the Internal Unadjusted data. There remains a pronounced rise in inequality based on the Internal-CC data between 1992 and 1993, but that rise is much smaller that the rise shown by the Internal-Unadjusted series between those years. This suggests that part of the 1993 jump is due to changes in censoring levels in the Internal-Unadjusted data in 1993. But even the smaller but pronounced increase in the Gini between 1992 and 1993 that remains in the Internal-CC series is unlikely to come from a genuine change in income inequality and suggests that other changes in CPS data collection methods are responsible. (See Ryscavage, 1995 and Jones and Weinberg, 2000.)

Because both the Public-MI and Internal-MI series capture the very highest incomes, it is not surprising that both lead to higher measured inequality levels than their respective unadjusted counterparts. Unlike the Public-Unadjusted series, the Public-MI series does not jump in 1995, suggesting that the jump in the Public-Unadjusted series at this point is purely due to introducing cell mean imputations into the public use data. On the other hand, the Gini estimates derived from the Internal MI series rose sharply in year 1993, showing the same pattern as Internal-Unadjusted series. This is additional evidence that the 1993 jump in Gini is primarily due to other survey design changes rather than data problems related to censoring.

Gini estimates derived from the Public-MI and Internal-MI series are relatively close to each other, suggesting that our assumptions about the parametric form of the income distribution work consistently despite the differences in censoring prevalence in the public use and internal CPS data. Before 1993, the two series are nearly identical. They diverged somewhat between the 1993 to 2004 period, but are still relatively close overall. The average of the Public-MI Gini coefficients since 1993 is 0.431, only slightly lower than the 0.439 average for the Internal-MI series.

We now use regression analysis to more formally investigate whether the differences in levels and trends in inequality differ between the data series. Table 2 reports the results of our analysis of differences between the Gini estimates derived from Public-CC, Internal-CC, Public-MI, Internal-MI and those derived from Internal-Unadjusted CPS data. We do not consider the Public-Unadjusted series as it contains less information than the internal CPS data and its artificial 1995 jump makes it flawed compared to the other alternatives.

The dependent variable (y) in the regressions is a Gini estimate for a given year and data series. The explanatory variables are: a constant, which is the Gini derived from the Internal-Unadjusted data; a time trend (t = 1, 2, …, 30), which summarizes the trend in this Gini; four dummy variables to represent other Gini data sources and thereby control for the difference between the Gini estimates from each series and the corresponding estimate from the Internal-Unadjusted series (pc = 1 if the series is Public-CC and 0 otherwise; ic = 1 if the series is Internal-CC and 0 otherwise; pm = 1 if the series is Public-MI and 0 otherwise; im = 1 if the series is Internal-MI and 0 otherwise); interactions between each of the series dummy variables (pc, ic, pm, and im) and time (t), which controls for the difference between the trends across series; a dummy variable for post-1992 years (u) which controls for a change in the levels after 1992 using the internal data; interactions between each of the Gini source variables (pc, ic, pm, and im) separately interacted with u, which controls for divergence in levels after 1992; an interaction between t and u, which controls for changes in the trend after

1992 using internal data; and the Gini source variables (*pc*, *ic*, *pm*, and *im*) separately interacted with *t* and *u*, which controls for divergence between trends after 1992.

First we discuss our results with respect to Gini estimates from the Internal-Unadjusted series. The significant positive coefficient for time trend (*t*) shown in Table 2, column 1, shows that inequality rose between 1975 and 1992. After 1992, although inequality continued to rise, the rate of increase slowed significantly compared to the pre-1993 period (note the coefficients on *t* and especially *t***u*). There was also a significant jump in inequality in the post-1992 years compared to pre-1993 years: observe the statistically significant positive coefficient for *u*. However, much of this increase in inequality results from changes in survey methods and internal censoring points, as we demonstrate by comparisons with the other Gini series.

The coefficients for *pc* and for interactions with it allow us to compare Gini estimates from the Public-CC series to the Internal-Unadjusted series. Because consistent censoring in the public use data imposes topcodes below those found in unadjusted public use data, it is unsurprising that the level of inequality is significantly lower using the Public-CC series than using the Internal-Unadjusted one. This can be seen for years prior to 1993 from the statistically significant negative coefficient for *pc* and, for years after 1992, from the statistically significant negative coefficient for interaction *pc***u*. Additionally, in 1993, when censoring points increased dramatically, we see that while inequality increased according to the Public-CC series (look at the coefficients on *u* + *pc***u*), the magnitude of the increase was substantially smaller than that in the Internal-Unadjusted series. This is also not surprising since consistent censoring dampens Gini levels in every year, and so is more likely to mitigate changes.

Despite the statistically significant difference in levels of inequality and the artificial jump in inequality in 1992, in years other than 1992–1993 the trends in inequality shown by the Internal-Unadjusted series are not significantly different from those shown by the Public-CC series: the coefficient on *pc***t* is not statistically significant. An F-test of the null hypothesis that *pc***t* + *pc***t***u*= 0 is not rejected, indicating that after 1992, the inequality trends in the two series are not significantly different.

Table 2
Differences in Inequality Levels and Trends Based on Regression Analyses

	Gini	*I*(0)	*I*(1)	*I*(2)
constant	0.3528**	0.2850**	0.2121**	0.2489**
	(0.0018)	(0.0048)	(0.0029)	(0.0082)
t (time trend)	0.0030**	0.0044**	0.0042**	0.0066**
	(0.0002)	(0.0004)	(0.0003)	(0.0008)
u (=1 if post-1992)	0.0592**	0.0190	0.1117**	0.3849**
	(0.0047)	(0.0261)	(0.0099)	(0.0454)
*t***u*	–0.0023**	0.0002	–0.0032**	–0.0071**
	(0.0002)	(0.0011)	(0.0005)	(0.0021)
pc (=1 if public CC)	–0.0158**	–0.0173*	–0.0250**	–0.0535**
	(0.0025)	(0.0073)	(0.0035)	(0.0085)
*pc***t*	–0.0003	–0.0005	–0.0009*	–0.0026**
	(0.0002)	(0.0006)	(0.0003)	(0.0008)
*pc***u*	–0.0282**	–0.0410	–0.0801**	–0.3586**
	(0.0067)	(0.0365)	(0.0119)	(0.0464)
*pc***t***u*	0.0006	0.0009	0.0015*	0.0059**
	(0.0003)	(0.0016)	(0.0006)	(0.0022)
ic (=1 if internal CC)	–0.0041	–0.0048	–0.0080*	–0.0204*
	(0.0026)	(0.0074)	(0.0037)	(0.0088)
*ic***t*	–0.0002	–0.0003	–0.0005	–0.0017*
	(0.0002)	(0.0006)	(0.0003)	(0.0008)
*ic***u*	–0.0176*	–0.0274	–0.0632**	–0.3338**
	(0.0067)	(0.0376)	(0.0126)	(0.0497)
*ic***t***u*	0.0004	0.0006	0.0012*	0.0061**
	(0.0003)	(0.0016)	(0.0006)	(0.0023)
pm (=1 if public MI)	0.0046*	0.0059	0.0115**	0.0382**
	(0.0023)	(0.0069)	(0.0038)	(0.0143)
*pm***t*	0.0001	0.0002	0.0006	0.0043**
	(0.0002)	(0.0006)	(0.0004)	(0.0015)
*pm***u*	–0.0188*	–0.0265	–0.0498*	–0.2143
	(0.0087)	(0.0361)	(0.0238)	(0.1565)
*pm***t***u*	0.0005	0.0007	0.0010	0.0055
	(0.0004)	(0.0015)	(0.0010)	(0.0065)
im (=1 if internal MI)	0.0043	0.0057	0.0122**	0.0508**
	(0.0026)	(0.0072)	(0.0040)	(0.0135)
*im***t*	0.0001	0.0001	0.0003	0.0029
	(0.0003)	(0.0006)	(0.0005)	(0.0018)
*im***u*	–0.0026	–0.0032	0.0002	0.1110
	(0.0073)	(0.0359)	(0.0256)	(0.2586)
*im***t***u*	0.0003	0.0005	0.0010	0.0114
	(0.0004)	(0.0015)	(0.0012)	(0.0115)

The baseline data source is Internal-Unadjusted. Numbers reported in parentheses are standard errors. ** Significant at the 1% level. * Significant at the 5% level. Source: Authors' calculations from internal and public use data files of March CPS.

We showed previously that the Gini estimates derived from Internal-Unadjusted data can be replicated using the Public-CM series. So the comparison just undertaken could be repeated with a comparison between inequality levels derived from the Public-CC and Public-CM series. But the Public-CC series is clearly dominated by the Internal-CC series because internal data allow us to observe the reported incomes of a greater share of the population, while still controlling for problems of time inconsistency in censoring.

Therefore, to check whether inequality trends of inequality according to Internal-Unadjusted data differ significantly from those found for lower parts of the income distribution that are consistently censored, we also examined the differences between the inequality series based on the Internal-CC and Internal-Unadjusted series. While the Internal-CC data imposes censoring points for observations at values below the incomes for the same observation in the Internal-Unadjusted data, the difference in levels of inequality is no longer significant for years prior to 1993: observe the negative but statistically insignificant coefficient for *ic*. But the difference for the years after 1993 is significantly lower as can be seen by the statistically significant negative coefficient for *ic* + *ic***u*. However, once again, the difference in inequality trends between the Internal-CC and Internal-Unadjusted series are not significantly different either before 1993 (observe the coefficient on *ic***t*) or after 1992 (*ic***t* + *ic***t***u*).

We now examine the levels and time trends shown by the two Gini series using our multiple imputation method. Because the multiple imputation method imputes income values for censored observations, both Public-MI and Internal-MI lead to Gini estimates that are larger than their Internal-Unadjusted counterparts, although the differences are only significant at the 5 percent level for the Public-MI series (note the coefficients for *pm* and *im*). The Public-MI series also jumps in 1993 (*u* + *pm***u*), but the increase is much less than the jump shown by the Internal-Unadjusted series (shown by a statistically significant negative coefficient for *pm***u*). In contrast, as shown by a small negative and insignificant coefficient for *im***u*, the Internal-MI series jumps in 1993 almost as much as the Internal-Unadjusted series. This again shows that changes in survey methods are important factors underlying the measured change in inequality levels after 1993. Our results are consistent with Census Bureau warnings that pre-1993 and post-1992 inequality measures based on data from CPS internal data are not directly comparable (Jones and Weinberg 2000).

Nevertheless, both Internal-MI and Public-MI series show the same linear time trend as the Internal-Unadjusted series, in both the pre-1993 and post-1992 periods. This is shown by the non-significance of the four time trend coefficients (*pm***t*, *pm***t***u*, *im***t*, and *im***t***u*).

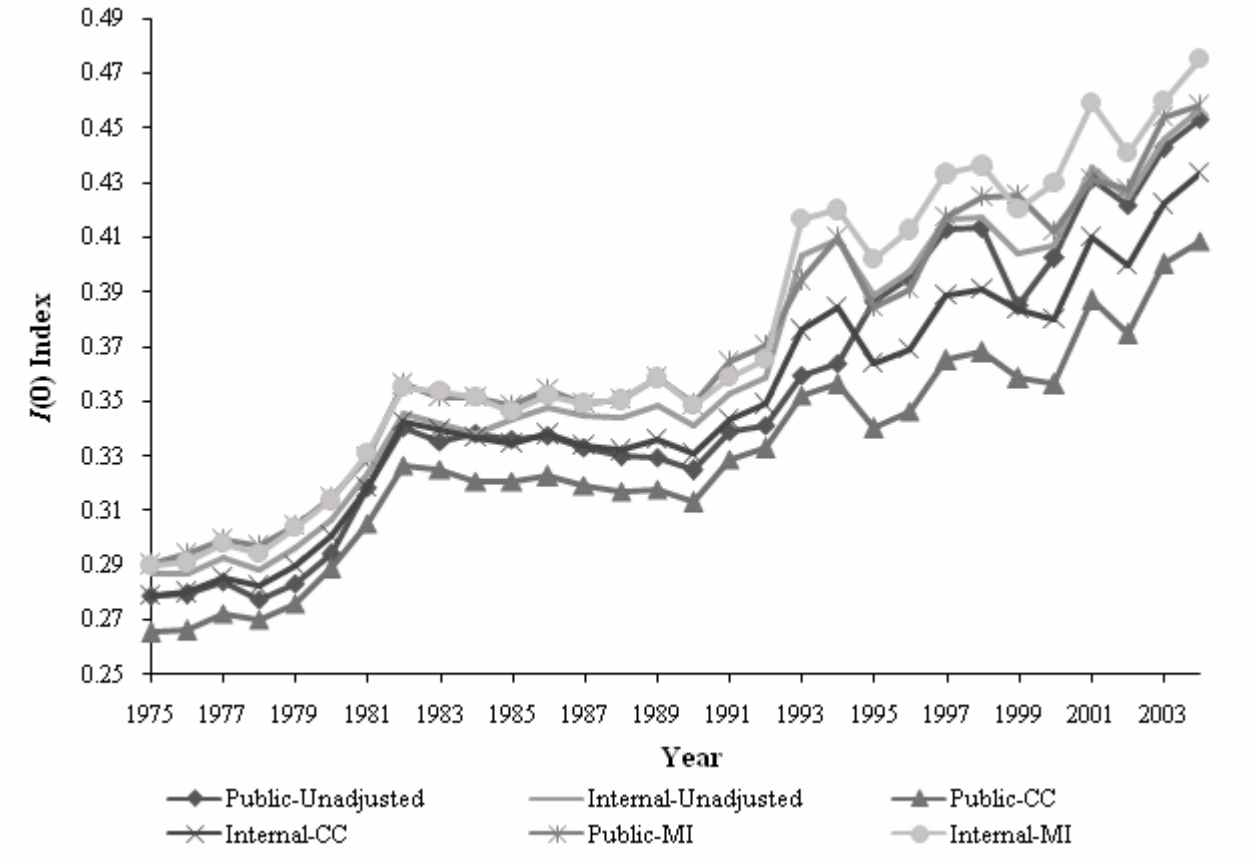
Thus, subject to the break in 1993, all five Gini series demonstrate the same time trends for the whole period 1975–2004. Also, the post-1992 rate of increase is much lower than the pre-1993 period.

Do results differ if different measures of inequality are used?

The Gini coefficient is just one of many commonly-used indices of inequality. It incorporates particular assumptions about how income differences are aggregated at different parts of the income distribution: it is relatively sensitive to income differences in the middle of the distribution ('middle sensitive'). This raises the question of whether alternative inequality measures, ones incorporating different aggregation assumptions, might lead to different conclusions about inequality trends, depending, for example, on whether changes in dispersion occur mainly at the top or the bottom of the distribution. We explore the sensitivity of our results to the choice of inequality measure by repeating our analysis using three indices from the Generalized Entropy parametric family. (See e.g. Cowell, 2000, or Jenkins and Van Kerm, 2009.) We employ the three most commonly used General Entropy (GE) measures: $I(0)$, the mean logarithmic deviation (MLD), which is a relatively bottom-sensitive index; $I(1)$, the Theil index, which is relatively middle-sensitive; and $I(2)$, half the squared coefficient of variation, which is an index that is top-sensitive.

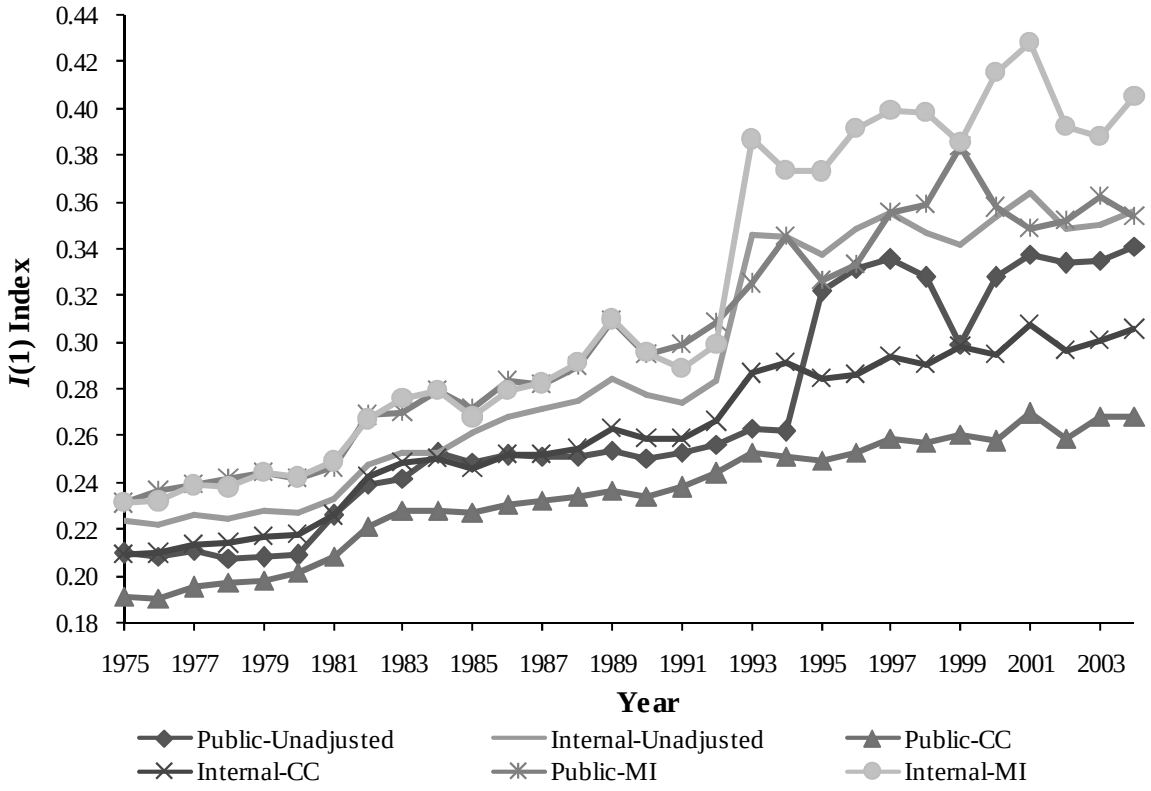
We calculated each of these GE indices using each of the six data series for which we derived Gini coefficients, and graph the estimates in Figures 4, 5 and 6 for $I(0)$, $I(1)$ and $I(2)$ respectively. In all cases, measured inequality in a given year is greater in an internal data series than in its public use data counterpart.

Figure 4
***I*(0) Inequality Index Estimates for Six Censoring Adjustment Methods, 1975–2004**



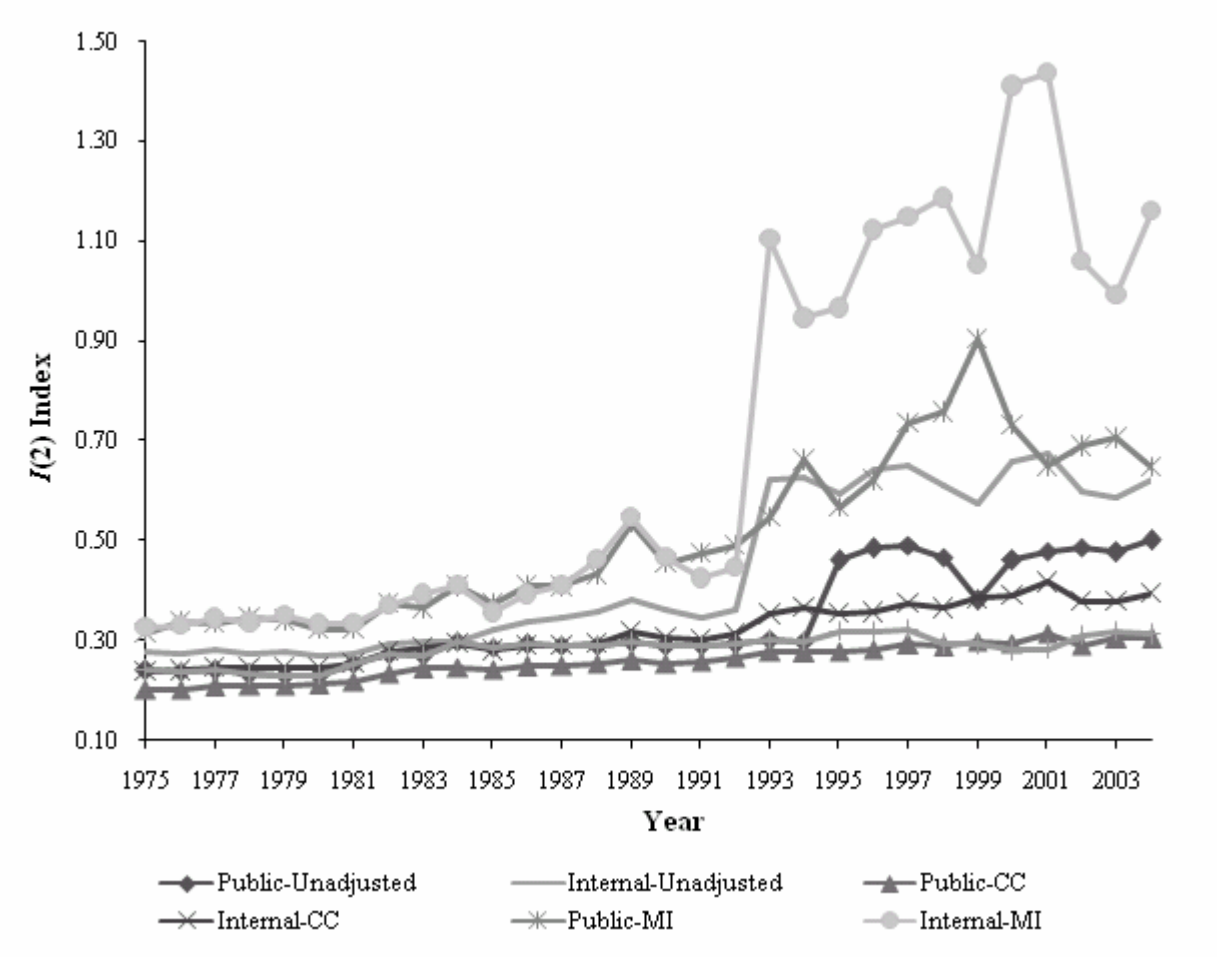
Source: Authors' calculations from internal and public use data files of March CPS.

Figure 5
***I*(1) Inequality Index Estimates for Six Censoring Adjustment Methods, 1975–2004**



Source: Authors' calculations from internal and public use data files of March CPS.

Figure 6
***I*(2) Inequality Index Estimates for Six Censoring Adjustment Methods, 1975–2004**



Source: Authors' calculations from internal and public use data files of March CPS.

Once again, the total period can be divided into three sub-periods, 1975–1992, 1992–1993, and 1993–2004. For each sub-period, we calculated the average annual percentage change in each GE index and the Gini according to the different series: see Table 3. Once again, the estimates based on Public-Unadjusted data are not shown because of the cell mean problem. Table 3 shows that the change in measured income inequality between 1992 and 1993 is much greater than the average change for each of the years before or after, regardless of the inequality index or data series used. This further suggests the changes between 1992 and 1993 represent changes in survey practice rather than a genuine change in the distribution of income.

Table 3 shows that average yearly increase in measured inequality is greater in the years prior to 1993 than in the years after 1993 for all measures across all data sets, but the difference is much less for the *I*(0) values than for any of the other measures. Because the *I*(0)

values put more weight on income differences in the lower ranges of the income distribution, this suggests that changes in the upper ranges of the income distribution might be driving the slowdown of the increase in inequality in the period 1993–2004.

Table 3
Average Annual Percentage Change in inequality by period, using alternative March CPS datasets and inequality indices

	Average annual percentage change		
	1975–1992	1992–1993	1993–2004
Gini Coefficients			
Unadjusted Internal	0.74	5.85	0.14
Consistent Censoring Public	0.75	1.93	0.23
Consistent Censoring Internal	0.74	3.24	0.21
MI-GB2 public	0.82	2.34	0.30
MI-GB2 Internal	0.77	6.63	0.20
$I(0)$			
Unadjusted Internal	1.48	12.48	1.20
Consistent Censoring Public	1.49	5.86	1.45
Consistent Censoring Internal	1.48	7.80	1.39
MI-GB2 public	1.63	6.41	1.47
MI-GB2 Internal	1.54	14.04	1.28
$I(1)$			
Unadjusted Internal	1.58	22.08	0.26
Consistent Censoring Public	1.63	3.67	0.55
Consistent Censoring Internal	1.61	7.71	0.59
MI-GB2 public	1.97	5.39	0.80
MI-GB2 Internal	1.72	29.32	0.43
$I(2)$			
Unadjusted Internal	1.76	71.82	0.03
Consistent Censoring Public	1.88	4.09	0.88
Consistent Censoring Internal	1.89	12.67	1.10
MI-GB2 public	3.31	11.22	1.69
MI-GB2 Internal	2.21	147.17	0.45

Source: Authors' calculations from internal and public use data files of March CPS.

As we did for our Gini estimates, we now use regression analysis to examine more precisely the differences in the levels and trends in inequality according to the different data series, where inequality is now measured by the three GE indices: see columns 2 through 4 of Table 2. The explanatory variables in the regressions are defined in an analogous fashion to those employed when comparing Gini levels and trends.

The results for $I(0)$ clearly differ from those based on the Gini measure: see Table 2, column 2. While the basic finding of a rise in inequality is still present for the Internal-

Unadjusted series – the coefficient on *t* is positive and statistically significant – there is no change in the index's trend after 1992, as the coefficient on *t***u* is not statistically significantly different from zero. All other series demonstrate the same time trend: the F-test that *t* + *pc***t*, *t* + *ic***t*, *t* + *pm***t*, and *t* + *im***t* each equal zero is soundly rejected at the 1 percent level in each case. Similarly, none of the other series show a slowdown in the rising trend for the period 1993–2004. Even at the 10 percent level, we cannot reject the null hypothesis that the coefficients on *t***u*+*pc***t***u*, *t***u*+*ic***t***u*, *t***u*+*pm***t***u*, and *t***u*+*im***t***u* equal zero.

With the exception that the level of inequality described by the Public-CC series is significantly lower than the corresponding values from Internal-Unadjusted series – the coefficient on *pc* is negative and significant – all other variables in this regression are statistically insignificant. In addition, the trends found in all five series are not significantly different over the entire period. Since *I*(0) is relatively bottom-sensitive, and the differences between these different series arise at the top of the distribution, it is not surprising that the choice of censoring method makes little difference for this index.

In contrast, the results for the *I*(1) values are very close to those found for the Gini coefficient: see Table 2, column 3. According to the Internal-Unadjusted series, the basic time trend (coefficient on *t*) is positive and statistically significant, as is the coefficient on *u*, while the coefficient on their interaction, *t***u*, is negative and statistically significant. Thus, as for the Gini index, inequality after 1992 according to *I*(1) is greater, but the annual increase in inequality since 1993 is significantly lower than in the years prior to 1993.

When the Public-CC and Internal-CC series are used rather than the Internal-Unadjusted one, the pre-1993 levels in *I*(1) are both significantly lower, as shown by the coefficients on *pc* and *ic*. For the post-1992 period, both series show significantly smaller jumps in levels: the coefficients on *pc***u* and *ic***u* are negative and statistically significant. But the post-1992 inequality levels are still significantly higher than those for the pre-1993 period, for both series. The null hypotheses that *u*+*pc***u* and *u*+*ic***u* are each equal to zero are both soundly rejected. With respect to trends, the coefficient on *pc***t* is negative and statistically significant, whereas the coefficient on *ic***t* is negative but not statistically significant.

The Public-MI and Internal-MI series both show higher levels of inequality than the Internal-Unadjusted series for corresponding years. The Public-MI series show a smaller jump in post-1993 levels, as shown by the negative and statistically significant coefficient on *pm***u*. Both series, though, show higher levels of inequality in the post-1993 period than in

the pre-1993 period. Both series show a clearly rising trend for the whole period, but the rates of increase in the post-1993 period are significantly smaller, as shown by F-tests on *t*u+pm*t*u* and *t*u+im*t*u*.

Finally, we turn to examine the levels and trends in *I*(2): see Table 2, column 4. According to the Internal-Unadjusted series, the basic time trend is positive and significant, as is the coefficient on *t*, while *t*u* is negative and significant. Therefore, as with the Gini and the *I*(1) indices, inequality after 1992 is greater but the annual increase in inequality since 1993 is significantly lower than the increase over the years prior to 1993.

When the Public-CC and Internal-CC series are compared to the Internal-Unadjusted one, the pre-1993 levels of *I*(2) are significantly lower for both series, as shown by the coefficients on *pc* and *ic*. For the post-1993 period, both series also show significantly smaller degrees of increase in levels because the coefficients of *pc*u* and *ic*u* are negative and statistically significant. But for both series, the post-1993 inequality levels are still significantly higher than the pre-1993 period. The null hypotheses that *u+pc*u* and *u+ic*u* each equal zero are both rejected. Both the CC series show smaller upward trend for the pre-1993 period than does the Internal-Unadjusted one: the coefficients on *pc*t* and *ic*t* are negative and statistically significant. Although the Public-CC series shows a lower trend after 1993 compared to before 1992, we cannot reject the null that the time trends before and after 1993 are the same in the Internal-CC series.

According to both the Public-MI and Internal-MI series, *I*(2) values are larger than the corresponding value in the Internal-Unadjusted series. In contrast to the time trend results for the Gini and *I*(1), the rates of increase in the post-1993 period are the same as the pre-1993 period. Using F-tests, we find that the coefficients on *t*u+pm*t*u* and *t*u+im*t*u* are not significantly different from zero. These results, as well as the contrast with those for the Gini and I(1), may arise because *I*(2) estimates tend to have larger sampling variability relative to these other indices, other things being equal.[5] This can also be seen from large year-on-year changes in *I*(2) values, and the large differences between their Public-MI and Internal-MI estimates, especially in the post-1993 period (see Figure 6). Thus, more caution is warranted when interpreting *I*(2) values.

[5] *I*(2) is also relatively prone to non-robustness to the effects of outliers in the sense discussed by Cowell and Victoria-Feser (1996). However, we would stress the sampling variability aspect, since the 'averaging' process used to combine the estimates from our 100 multiply imputed data sets are likely to smooth out the effects of any outliers being added by the imputation process.

Trends in Top Income Shares

Our results may surprise some readers – occasional users of the CPS may not be aware of the problems that inconsistent censoring can cause. For example, they may not be aware of comparability issues introduced by the Census Bureau's use of cell mean imputations beginning in 1995, and the changes in the CPS (public use and internal data) before and after 1993.[6] Other readers may be surprised by our findings because they appear to contradict the seminal work of Piketty and Saez (2003), who report substantial growth in income inequality as measured by the share of "adjusted gross income" (the amount of income that is taxable under federal income tax laws in place each year) held by tax units (unadjusted for household size) at the top of the distribution.

Up to now, we have focused on how best to use CPS data to make statements about levels and trends in income inequality as it has conventionally been measured in the income distribution literature. (See Gottschalk and Smeeding, 1997, and Atkinson and Brandolini, 2001, for reviews.) In this section, we compare our CPS-based estimates of levels and trends in the income distribution with those of Piketty and Saez (2003).[7] In doing so, it is critical to point out that there are fundamental differences between the measurement conventions used when analyzing size-adjusted pre-tax post-transfer household income of individuals using cross-sectional surveys like the CPS and using IRS data on the adjusted gross income of tax units as done by Piketty and Saez (2003). Piketty and Saez (2003) were primarily interested in measuring very long run trends (since 1913) in US income inequality and, because of data limitations, focus only on a small part of that distribution – the share of adjusted gross income held by the top 10 percent of tax filing units and how that share is distributed within this part of the distribution.

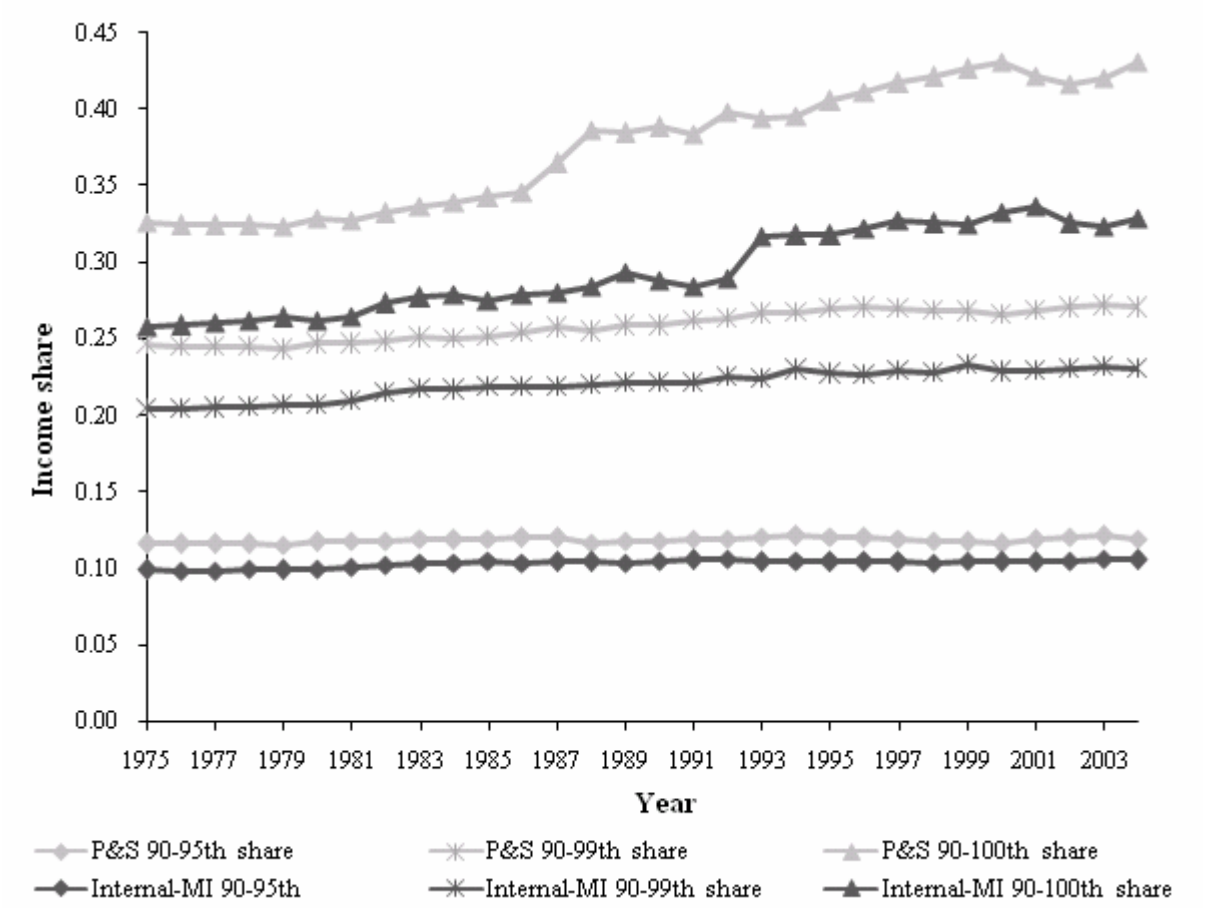
In Figure 7 we report our estimates of the share of total income (as discussed in the first part of this paper) held by the richest 10 percent of the population. We do so by cumulatively reporting the share held by the 90^{th}-95^{th} percentile group and then by this group

[6] For instance, the most important source of standardized cross-sectional micro data on industrialized countries—the Luxembourg Income Study (LIS)—uses the public use version of the CPS for its data on the USA. On its website (http://www.lisproject.org/keyfigures/ineqtable.htm), LIS summarizes income inequality using P90/P10 and Gini coefficient estimates that do not adjust for the topcoding issues discussed here. Nor does LIS currently provide its users with information on the degree that topcoding issues of this type are found in its data from other countries. The public use CPS data as developed by LIS is also a major source of information about USA inequality in the United Nation University—World Income Inequality Database (UNU-WIDER, 2008). UNU-WIDER also does not alert its users of topcoding issues on its website (http://www.wider.unu.edu/research/Database/en GB/database).

[7] Piketty and Saez (2003) only constructed top income shares up to year 1998, but they subsequently updated this series to year 2006, see http://elsa.berkeley.edu/~saez/. It is these latter values that we report here.

plus the next richest 4 percent (95th-99th percentile group) and finally by adding on the share held by the top 1 percentile group for each year 1975–2004, based on the Internal-MI series. We also report the corresponding cumulative share estimates from Piketty and Saez (2003) for their tax unit samples and use of adjusted gross income (excluding capital gains).

Figure 7
Estimates of Top Income Shares: Piketty & Saez and Internal-MI Series, 1975–2004



Source: http://elsa.berkeley.edu/~saez/ and authors' calculations from internal data files of March CPS.

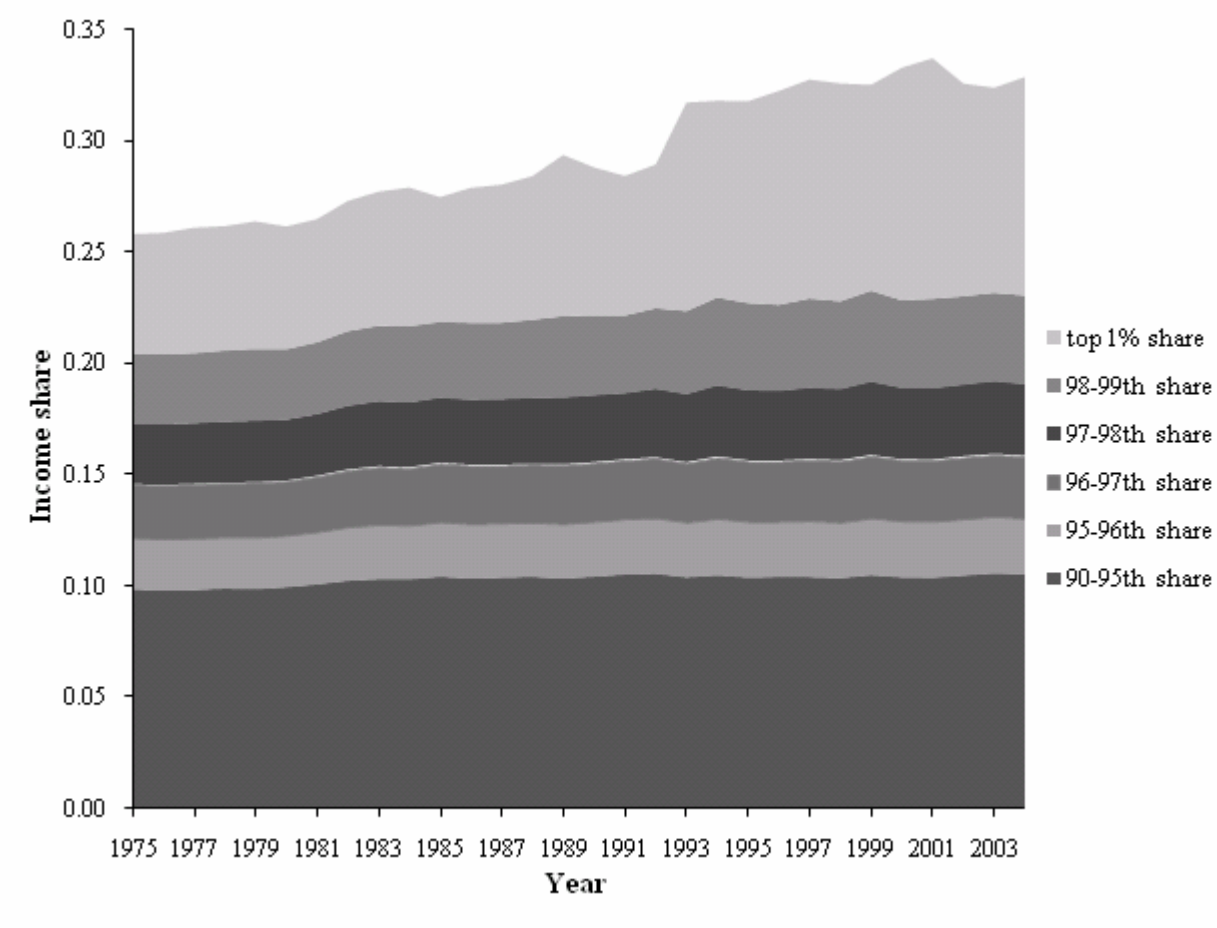
Our Internal-MI top share estimates are in general somewhat lower than the corresponding ones reported by Piketty and Saez (2003). This is not surprising since two quite different concepts of income are being used. For instance, adjusted gross income does not include government transfers since this income is not taxed. Because a much greater share of non-taxable government in-cash transfers – AFDC/TANF, Social Security benefits, etc. – are held by those in the poorest 90 percent of the pre-tax post transfer distribution, and this income is included in our income measure, we would expect those in the top 10 percent

of the pre-tax post-transfer income distribution to have a smaller share of total income than of adjusted gross income in all years, as is the case in Figure 7.

Nevertheless, for both the income share held by the 90th–95th percentile groups and the 95th–99th percentile groups, the CPS-based and tax data-based series resemble each other closely, in terms of both levels and trends. This is remarkable given the very different underlying concepts of income and approaches taken. Both the Piketty and Saez (2003) IRS series and our Internal-MI CPS series suggest that the share of income held by the 90th–95th percentile groups remained relatively stable during the whole period 1975–2004, accounting for around 10–12 percent of total income. In terms of the share of income held by the 95th–99th percentile groups, the Piketty and Saez (2003) IRS-based estimate grows slowly from 13 percent in 1975 to 15.1 percent in 2004. Over the same period, the growth in our corresponding Internal-MI CPS estimate is remarkably similar: an increase from 10.5 percent to 12.5 percent.

In Figure 8, we disaggregate the share of income held by the richest 10 percent of the income distribution further by reporting the share of income held by percentile groups within the 95th–99th percentile group according to our Internal-MI CPS series. Although income shares increased slightly at the higher percentiles, through to the 99th percentile group, these increases were relatively modest. Even for the 98th–99th percentile group, the growth in income shares from 1975–2004 was less than 1 percentage point.

Figure 8
Estimates of Top Income Shares: Internal-MI Series, 1975–2004



Source: Author's calculations using internal March CPS data

In contrast to the similar trends found in the share of income held by those in the 90th–99th percentile group using either the Piketty and Saez (2003) series or our Internal-MI series, there are substantial differences between the two series with respect to the level and trends in the share of income held by the richest 1 percent. While the share of income help by the richest 1 percent grew substantially over the 1975–2004 period according to both series, growth in the Piketty and Saez (2003) series is much greater, from 8.0 percent to 16.1 percent compared to an increase from 5.4 percent to 9.8 percent in the Internal-MI CPS series. But this difference is even greater when it is observed that one third of the increase in the income share of the top 1 percent in the Internal-MI series from 1975–2004 occurred in 1993, and which is primarily attributable to the CPS redesign. Hence a significant minority of this increase in the share of income held by the top 1 percent in the Internal-MI series is likely to be due to better measurement of their income by the CPS rather than by an actual increase in the share of income they held.

This same problem of changes in measurement versus changes in real income held by

the richest 1 percent of the distribution is also likely to explain at least some part of the rise in income shares at the top of the income distribution in the Piketty and Saez (2003) series. As can be seen in Figure 7, there is a dramatic 4 percentage point jump in the share of gross taxable income held by the highest 1 percent of tax units reported by Piketty and Saez (2003) between the years 1986 and 1988. This finding must, to some degree, be the result of changes in the way the very richest tax units chose to report their income as a result of the Tax Reform Act of 1986 rather than the result of genuine changes in income inequality. This legislation provided substantial incentives for the very richest tax units to switch income from Subchapter-S corporations to wage income. (See Slemrod, 1996, and Reynolds, 2006 for a fuller discussion of this type of tax elasticity issue in the Piketty and Saez, 2003 data and Feenberg and Poterba, 1993 for a more general discussion of the problems of measuring income inequality based on income tax records.) Subtracting the 1986–1988 jump in the Piketty and Saez series would cut in half the increase in the share of income held by the richest 1 percent over the whole period. Their data for other years may also be subject to this same type of tax elasticity behavior, albeit to a lesser extent. For example, for a group of high income corporate executives, Goolsbee (2000) showed that almost all of the estimated responsiveness of taxable wage and salary income to marginal tax rates from 1991 to 1995 was the result of shifts in the timing of compensation rather than permanent shifts in the form of compensation.

Keeping this in mind, it is difficult to determine what is going on at the very top of the income distribution using either the CPS data or the IRS data. Piketty and Saez (2003) find greater increases in inequality after 1993 than we do primarily because of greater growth in the share of income held by the richest 1 percent in their IRS data. It is uncertain to what degree this difference is the result of our decreasing ability to capture income at the very highest income levels, even using internal CPS data, or of behavioral changes in the way that individual tax units report their adjusted gross income on their tax returns.

The growth in income in the years just prior to the 2000 stock market crash was greater among the highest income groups, and the estimates of the income share of the richest 1 percent from both series capture this increase as well as the decline in their share in the recession that followed the stock market crash. The CPS may have been less able to capture this income going to the top of the income distribution. But it also may be the case, as Reynolds (2006) argues, that a greater increase in the use of tax-deferred savings accounts (401k plans, Keogh plans and IRA tax shelters) by those in richer percentile groups but not in

the very richest 1 percent of the adjusted gross income distribution of tax units may also explain part of the rise in the top income share reported by Piketty and Saez (2003).

Additional work is necessary to determine what precisely is happening at the very highest income levels. But one more piece of evidence with respect to the poorest 99 percent of the distribution – which can be measured with confidence in the CPS data – is that our trends for income inequality are consistent with the trends in wage earnings over this period found by Autor, Katz and Kearney (2008) using the public use CPS data.

Conclusions

We analyze trends in income inequality in the USA over three decades (1975–2004), drawing on internal CPS data that are not generally available to researchers outside the Census Bureau. Although the CPS is the most authoritative survey data source for studying US income inequality, censoring of each of its many sources of income artificially lowers the level of inequality and potentially affects estimates of its trends over time, whether one uses public or internal CPS data. We consider several methods for addressing censoring, and as a byproduct of this work, derive a revised cell mean imputation series that can be applied to public use data. This series allows researchers to more accurately capture levels and trends in the internal data. In addition, we take account of censoring in the internal data using a multiple imputation approach that uses parametric methods to derive imputations.

Using the Gini coefficient to measure inequality, we find the level of income inequality based on our Internal-MI series is slightly higher in corresponding years inequality based on the Internal-Unadjusted series before 1993 and significantly higher after 1993. But despite this difference in levels, the *trends* in inequality we find using the two series are not significantly different. Inequality rose over the whole period 1975–2004, but the rate of increase slowed after 1993. We find this same result using our Internal-CC and Public-CC estimates.

The story about trends differs when we use the three GE inequality indices, but the differences are readily explicable in terms of the properties of the indices, specifically their differences in sensitivity to income differences in different ranges of the income distribution. $I(0)$ rose steadily during 1975–2004, showing no sign of slowing down after 1993. But our $I(1)$ and $I(2)$ series are similar to our Gini series in their trends, subject to the caveat that $I(2)$ indices are less precisely estimated, showing large year-on-year variation. Because $I(0)$ is more bottom-sensitive than the other inequality measures, our results suggest that what is

happening in the upper part of the income distribution is responsible for the 1990s slowdown in rate of inequality increase.

Our results for household income are largely consistent with other research that has examined levels and trends in wage inequality using public use CPS data (see e.g. Autor, Katz and Kearney, 2008) but differ from the results found by Piketty and Saez (2003) after 1986 almost solely due to different trends in the share of income held by the richest 1 percent. With IRS data about the distribution of adjusted gross income, not only do Piketty and Saez find higher levels of inequality over all years than we do, but also a major increase in inequality between 1986 and 1988 and a greater increase in income inequality trends in the post-1993 period. Although the rapid increase in income inequality in their tax data is caused in part by changes in the way income is reported rather than a real change in underlying income inequality, it is less certain why they find such larger increased in income inequality after 1993.

Our results suggest that, for at least the poorest 99 percent of the income distribution, the increase in inequality since 1993 has been significantly slower in the USA than in the previous two decades. Based on our Internal-MI series we find the level of income inequality rises when we include an estimate that includes the very top part of the income distribution censored in the unadjusted internal CPS data. But even in these estimates, the rise in income inequality slowed after 1993. Our findings are consistent with those found by Piketty and Saez (2003) for the 90^{th}–99^{th} percentile groups. It is only with respect to the richest 1 percent that we differ. And it is here that we are at the limits of current knowledge, both with respect to the CPS data because of its difficulty in obtaining information on the highest income households, and with respect to the IRS data because of behavioral effects caused by changes in the tax laws. It is difficult to fully understand how much of the yearly changes in inequality are the result of real changes in the incomes of the very richest income tax units and how much is due simply to changes in the way they report that income.

References

Atkinson, Anthony B., and Andrea Brandolini. 2001. “Promises and Pitfalls in the Use of Secondary Data Sets: Income Inequality in OECD Countries as a Case Study.” *Journal of Economic Literature,* 39 (3): 771–799.

Atkinson, Anthony B., Lee Rainwater, and Timothy Smeeding. 1995. *Income Distribution in OECD Countries. Evidence from the Luxembourg Income Study*. Social Policy Studies No. 18. Paris: Organization for Economic Cooperation and Development.

Autor, David, Lawrence Katz, and Melissa Kearney. 2008. "Trends in U.S. Wage Inequality: Revising the Revisionists." *Review of Economics and Statistics* 90(2): 300–323.

Bandourian, Ripsy, James B. McDonald, and Robert S. Turley. 2002. "A Comparison of Parametric Models of Income Distribution across Countries and Over Time." Working Paper No. 305. Luxembourg: Luxembourg Income Study. http://www.lisproject.org/publications/LISwps/305.pdf

Bishop, John A., Jong-Rong Chiou, and John P. Formby. 1994."Truncation Bias and the Ordinal Evaluation of Income Inequality." *Journal of Business and Economic Statistics*, 12: 123–127.

Bordley, Robert F., James B. McDonald, and Anand Mantrala. 1996. "Something New, Something Old: Parametric Models for the Size Distribution of Income." *Journal of Income Distribution* 6: 91–103.

Brachmann, Klaus, Andreas Stich, and Mark Trede. 1996. "Evaluating Parametric Income Distribution Models." *Allgemeines Statistisches Archiv* 80: 285–298.

Burkhauser, Richard V., J.S. Butler, Shuaizhang Feng, and Andrew Houtenville. 2004. "Long-Term Trends in Earnings Inequality: What the CPS Can Tell Us." *Economics Letters*, 82: 295–299.

Burkhauser, Richard V., Kenneth A. Couch, Andrew J. Houtenville, and Ludmila Rovba. 2003-2004. "Income Inequality in the 1990s: Re-Forging a Lost Relationship?" *Journal of Income Distribution,* 12 (3–4): 8–35.

Burkhauser, Richard V., Shuaizhang Feng, and Stephen P. Jenkins. 2007. "Using the P90/P10 Ratio to Measure US Inequality Trends with Current Population Survey Data: A View from Inside the Census Bureau Vaults." ISER Working Paper 2007-14. Colchester, UK: Institute for Social and Economic Research, University of Essex. http://www.iser.essex.ac.uk/pubs/workpaps/pdf/2007-14.pdf

Burkhauser, Richard V. and Jeff Larrimore. 2008. "Trends in the Relative Economic Well Being of Working-Age Men with Disabilities: Correcting the Record using Internal Current Population Survey Data." Center for Economic Studies Working Paper Series CES-WP-08-05, March 2008. http://www.ces.census.gov/index.php/ces/cespapers?down_key=101811

Burkhauser, Richard V., Takashi Oshio, and Ludmila Rovba. 2008. “How the Distribution of After-Tax Income Changed Over the 1990s Business Cycle: A Comparison of the United States, Great Britain, Germany, and Japan.” *Journal of Income Distribution,* 17 (1): 87–109.

Cowell, Frank A. 2000. “Measurement of Inequality.” In: Anthony B. Atkinson and François Bourguignon (eds), *Handbook of Income Distribution*, 87–166. Elsevier North Holland: Amsterdam.

Cowell, Frank A., and Maria-Pia Victoria-Feser. 1996. “Robustness Properties of Inequality measures.” *Econometrica*, 64: 77–101.

Daly, Mary C. and Robert G. Valletta. 2006. “Inequality and Poverty in United States: The Effects of Rising Dispersion of Men’s Earnings and Changing Family Behaviour.” *Economica*, 73 (289): 75–98.

Danziger, Sheldon and Peter Gottschalk. (Eds.) 1993. *Uneven Tides: Rising Inequality in America*. New York: Russell Sage Foundation.

Feenberg, Daniel and James M. Poterba. 1993. “Income Inequality and the Incomes of Very High-Income Taxpayers: Evidence from Tax Returns,” in *Tax Policy and the Economy*, ed. James M. Poterba, 7, Cambridge, MA: NBER/MIT Press, pp.

Feng, Shuaizhang, Richard V. Burkhauser, and J.S. Butler. 2006. “Levels and Long-Term Trends in Earnings Inequality: Overcoming Current Population Survey Censoring Problems Using the GB2 Distribution.” *Journal of Business and Economic Statistics* 24 (1): 57–62.

Fichtenbaum, R. and Shahidi, H. 1988. “Truncation Bias and the Measurement of Income Inequality.” *Journal of Business and Economic Statistics*, 6: 335–337.

Goolsbee, Austan. 2000. “What Happens When You Tax the Rich? Evidence from Executive Compensation.” *Journal of Political Economy*, 108 (2): 352–378.

Gottschalk, Peter, and Timothy M. Smeeding. 1997. “Cross-National Comparisons of Earnings and Income Inequality.” *Journal of Economic Literature,* 35 (2): 633–687.

Gottschalk, Peter, and Sheldon Danziger. 2005. “Inequality of Wage Rates, Earnings and Family Income in the United States, 1975–2002.” *Review of Income and Wealth* 51 (2): 231–254.

Jenkins, Stephen P., and Philippe Van Kerm. Forthcoming. “The Measurement of Economic Inequality.” In: W. Salverda, B. Nolan and Timothy M. Smeeding (eds), *The Oxford Handbook on Economic Inequality*. Oxford: Oxford University Press.

Jenkins, Stephen P., Richard V. Burkhauser, Shuaizhang Feng, and Jeff Larrimore. 2008. "Measuring Inequality with Censored Data: A Multiple Imputation Approach." ISER Working Paper. Colchester, UK: Institute for Social and Economic Research, University of Essex. (in preparation).

Jones, A.F., Jr., and Weinberg, D.H. 2000. "The Changing Shape of the Nation's Income Distribution." *Current Population Reports*, U.S. Census Bureau, June 2000. http://www.census.gov/prod/2000pubs/p60-204.pdf

Karoly, Lynn A. and Gary Burtless. 1995. "Demographic Changes, Rising Earnings Inequality, and the Distribution of Personal Well-Being, 1959–1989." *Demography*, 32 (3): 379–405.

Kruger, Dirk, and Fabrizio Perri. 2006. "Does Income Inequality Lead to Consumption Inequality?" *Review of Economic Studies*, 73 (1): 163–193.

Larrimore, Jeff, Richard V. Burkhauser, Shuaizhang Feng and Laura Zayatz. Forthcoming. "Consistent Cell Means for Topcoded Incomes in the Public Use March CPS (1976–2007)." *Journal of Economic and Social Measurement*, NBER working paper w13941

Levy, Frank, and Richard J. Murnane. 1992. "U.S. Earnings Levels and Earnings Inequality: A Review of Recent Trends and Proposed Explanations." *Journal of Economic Literature*, 30 (3): 1333–1381.

Luxemburg Income Study. 2008. Income Inequality Measures. http://www.lisproject.org/keyfigures/ineqtable.htm accessed August 1, 2008.

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

Reiter, J.P. 2003. "Inference for Partially Synthetic, Public Use Microdata Sets." *Survey Methodology*, 29: 181–188.

Reynolds, Alan. 2006. *Income and Wealth*. Westport, Connecticut: Greenwood Press.

Rubin, Donald B. 1987. *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley.

Ryscavage P. 1995. "A surge in growing income inequality?" *Monthly Labor Review*, August, 51–61. http://www.bls.gov/opub/mlr/1995/08/art5full.pdf

Slemrod, Joel. 1996. "High-income Families and the Tax Changes of the 1980s: The Anatomy of Behavioral Response", In *Empirical Foundations of Household Taxation*, edited by Martin Feldstein and James M. Poterba. Chicago: University of Chicago Press (for NBER).

UNU-WIDER, 2008. *World Income Inequality Database, version 2.0c (May).* United Nations University – World Institute for Development Economics Research, Helsinki.

http://www.wider.unu.edu/research/Database/en_GB/database accessed August 1, 2008.

U.S. Census Bureau. Various years. *Income, Poverty, and Health Insurance Coverage in the United States: various years*. (August), Current Population Reports, Consumer Income, Washington DC: GPO.

U.S. Census Bureau. 2007. Current Population Survey Annual Social and Economic Supplement Technical Documentation. Washington DC: GPO.

Welch, Finis. 1999. "In Defense of Inequality" *American Economic Review* 89: 1–17.

Welniak, Edward J. 2003. "Measuring Household Income Inequality Using the CPS." in James Dalton and Beth Kilss (Eds.), *Special Studies in Federal Tax Statistics 2003*, Statistics of Income Directorate, Inland Revenue Service, Washington DC.

Appendix Table 1. Income Items Reported in the Current Population Survey

Name	Name in Public Files	Name in Internal Files	Definition
		1975–1986	
Labor Earnings			
Wages	I51A	WSAL_VAL	Wages and Salaries
Self Employment	I51B	SEMP_VAL	Self employment income
Farm	I51C	FRSE_VAL	Farm income
Other Sources			
Social Security	I52A	I52A_VAL	Income from Social Security and/or Railroad Retirement
Supplemental Security	I52B	SSI_VAL	Supplemental Security Income
Public Assistance	I53A	PAW_VAL	Public Assistance
Interest	I53B	INT_VAL	Interest
Dividends Rentals	I53C	I53C_VAL	Dividends, Rentals, Trust Income
Veterans	I53D	I53D_VAL	Veteran's, unemployment, worker's compensation
Retirement	I53E	I53E_VAL	Pension Income
Other	I53F	I53F_VAL	Alimony, Child Support, Other income
		1987–2006	
Labor Earnings			
Primary earnings	ERN_VAL	ERN_VAL	Primary Earnings
Wages	WS_VAL	WS_VAL	Wages and Salaries-Second Source
Self Employment	SE_VAL	SE_VAL	Self employment income -Second Source
Farm	FRM_VAL	FRM_VAL	Farm income -Second Source
Other Sources			
Social Security	SS_VAL	SS_VAL	Social Security Income
Supplemental Security	SSI_VAL	SSI_VAL	Supplemental Security Income
Public Assistance	PAW_VAL	PAW_VAL	Public Assistance & Welfare Income
Interest	INT_VAL	INT_VAL	Interest
Dividends	DIV_VAL	DIV_VAL	Dividends
Rental	RNT_VAL	RNT_VAL	Rental income
Alimony	ALM_VAL	ALM_VAL	Alimony income
Child Support	CSP_VAL	CSP_VAL	Child Support Income
Unemployment	UC_VAL	UC_VAL	Unemployment income
Workers Comp	WC_VAL	WC_VAL	Worker's compensation income
Veterans	VET_VAL	VET_VAL	Veteran's Benefits
Retirement - Source 1	RET_VAL1	RET_VAL1	Retirement income - source 1
Retirement - Source 2	RET_VAL2	RET_VAL2	Retirement income - source 2
Survivors - Source 1	SUR_VAL1	SUR_VAL1	Survivor's income - source 1
Survivors - Source 2	SUR_VAL2	SUR_VAL2	Survivor's income - source 2
Disability - Source 1	DIS_VAL1	DIS_VAL1	Disability income - source 1
Disability - Source 2	DIS_VAL2	DIS_VAL2	Disability income - source 2
Education assistance	ED_VAL	ED_VAL	Education assistance
Financial assistance	FIN_VAL	FIN_VAL	Financial Assistance
Other	OI_VAL	OI_VAL	Other income

Sources: Current Population Survey Annual Demographic File Technical Documentation, 1976-2002. Current Population Survey Annual Social and Economic Supplement Technical Documentation, 2003-2007

Appendix Table 2: Rate of censoring when using consistent censoring procedures

	Public Consistent Censoring Points		Internal Consistent Censoring Points	
	Percent Censored	Binding Year	Percent Censored	Binding Year
Wage	1.31	2001	0.36	1984
Self Employment	3.93	1978	1.38	1984
Farm	2.42	2001	0.28	1983
Public Assistance	0.48	1995	0.48	1995
Supplemental Security	1.94	1994	0.30	1992
Social Security	0.40	1997	0.40	1997
Interest	1.15	2004	0.21	2004
Dividends	1.58	2001	0.15	1984
Veterans / Workers Comp	0.23	1985	0.00	1986
Retirement	0.17	1979	0.10	1984
Other Income	0.17	1979	0.07	1984

Source: Author's calculations using public use and internal March CPS data