Measuring Levels and Trends in Earnings Inequality with Nonresponse, Imputations, and Topcoding*

Christopher R. Bollinger, University of Kentucky
Barry T. Hirsch, Georgia State University and IZA, Bonn
Charles Hokayem, Centre College
James P. Ziliak, University of Kentucky

This version: December 2015

Abstract

Measures of U.S. earnings (and income) inequality rely heavily on the Current Population Survey Annual Social and Economic Supplement (ASEC). A substantial and increasing share of individuals and households surveyed in the CPS either do not participate in the ASEC supplement, or participate but fail to report earnings. Imputation procedures assume that nonresponse is missing at random, conditional on measured covariates (MAR). Yet little is known how deviations from MAR affect inequality measures. We explore how nonresponse bias affects measures of the level and trends in earnings inequality using ASEC data for calendar years 1997-2010 matched to Social Security Detailed Earnings administrative tax records (DER). We find that nonresponse bias causes inequality to be understated, with ASEC earnings responses including too few low earners and too few very high earners. Census hot-deck imputations for nonrespondents do not fully correct the bias. Earnings shares among the top 1% of earners are lower by at least 20 percent in the ASEC compared to matched administrative tax records, with about half accounted for nonresponse and half to topcoding in the ASEC. Hybrid measures using ASEC earnings for CPS respondents and administrative DER earnings for nonrespondents produce intermediate estimates.

JEL code: J31, Wage level and structure

Keywords: earnings inequality, nonresponse bias, imputations, topcodes, CPS-ASEC, administrative data, validation studies

* We thank seminar participants at the 4th SOLE/EALE World Conference, the 2015 Southern Economic Association Meetings, the 2nd Annual Federal Reserve Bank of Cleveland and University of Kentucky Workshop, the Russell Sage Foundation, and Columbia University for helpful comments on earlier versions. We are grateful to the U.S. Census Bureau for providing access to the restricted data used in this study. The views and opinions are solely those of the authors and do not reflect the views of the Census Bureau or any other sponsoring agency.

Address correspondence to: James P. Ziliak, Department of Economics, University of Kentucky, Lexington, KY 40506-0034; Email: jziliak@uky.edu

In the literature examining the levels and trends in earnings and income inequality in the United States, the principal data source has been public use files from the Current Population Survey (CPS). These studies have typically used either the CPS Outgoing Rotation Group (ORG) monthly files that measure earnings on the primary job during the survey reference week or the CPS Annual Social and Economic Supplement (ASEC) administered each March, providing information on earnings and income sources during the previous calendar year, or both (e.g., Bound and Johnson 1992; Juhn, Murphy, and Pierce 1993; DiNardo, Fortin, and Lemieux 1996; Card and DiNardo 2002; Lemieux 2006; Autor, Katz, and Kearney 2008). Burkhauser et al. (2012) have authored a series of papers examining trends in income inequality using *internal* ASEC files that have the advantage of earnings topcodes extending well beyond those in the public use files. Less common have been studies of inequality using administrative tax data (Piketty and Saez 2003; Kopczuk, Saez, and Song 2010; Chetty et al. 2014; Spletzer 2014). In this paper, we use both—internal ASEC data that is matched to administrative tax data on earnings—to estimate levels and trends in earnings inequality.

A complicating factor in the analysis of earnings inequality has been the substantial increase in earnings nonresponse in the CPS (Hirsch and Schumacher 2004; Bollinger and Hirsch 2013; Hokayem, Bollinger, and Ziliak 2015). Today over 30 percent of earnings values in the ASEC and ORG are imputed (allocated) based on the earnings of "donors" matched to nonrespondents in Census hot deck procedures, which is double the rate of two decades ago. Depending on the questions being addressed and the reasons for nonresponse, use of imputed values can either have little effect or can produce severe "match bias" in standard measures of wage gaps (Hirsch and Schumacher 2004; Bollinger and Hirsch 2006, Heckman and LaFontaine 2006). If the earnings data are "missing completely at random" (MCAR), then nonresponse is not dependent on earnings, even absent covariates; if earnings are "missing at random" (MAR), then nonresponse is not dependent on earnings after conditioning on covariates; and if earnings are "not missing at random" (NMAR), then nonresponse depends on the value of missing earnings even after conditioning on covariates (Rubin 1976; Little and Rubin 2002). It is this latter case that is generally referred to as "nonresponse bias." Both Census imputation procedures and common methods to deal with nonresponse assume that nonresponse is ignorable; that is, those not reporting earnings have earnings similar to those with equivalent measured attributes. If this MAR assumption is violated, then measures of inequality will be biased. Indeed, some inequality

studies have excluded imputed earners (Lemieux 2006; Autor, Katz, and Kearney 2008), while others have not (e.g., Burkhauser et al. 2012).

Although the literature has made progress in understanding how treatment of imputed earners affects wage equation coefficients and measures of wage gaps across groups (Bollinger and Hirsch 2006; Kline and Santos 2013; Bollinger et al. 2014), there is little knowledge regarding how earnings nonresponse affects the measurement of inequality. Understanding how nonresponse might affect inequality is not straightforward. One needs to identify who fails to respond, how nonresponse differs with respect to true and typically unobserved earnings (conditional on covariates), how any such nonresponse bias might differ across the earnings distribution, and how one can best treat topcoded earnings. Census uses different topcode values depending on earnings source, and these limits differ between internal and public release versions of the ASEC.

To address these questions, we use restricted-access ASEC data for calendar years 1997 through 2010, matched to administrative tax data on wage and salary and self-employment earnings from the Social Security Detailed Earnings Records (DER). A key advantage of the DER data is that earnings are not topcoded, thus permitting a direct comparison of estimates of upper-tail inequality from tax records to topcoded survey responses. This is the first such direct comparison from matched individual survey and tax data on how nonresponse and topcoding affects earnings inequality estimates. We estimate several leading measures of inequality emphasized in the recent literature—including the Gini coefficient, the earnings share accruing to the top 1% of earners, and the 90/50 and 50/10 percentile ratios—for alternative earnings definitions. The latter include the internal ASEC data inclusive of Census hot-deck imputations for nonrespondents, ASEC with nonrespondents dropped, and replacement of ASEC earnings with DER earnings. We also explore possible "fixes" for nonresponse in publicly available ASEC data, including re-weighting techniques for nonresponse and replacement of topcoded values in the ASEC with newly estimated values based on the Pareto distribution methods proposed by Armour, et al. (2014).

Our results indicate that nonresponse bias causes inequality to be understated, with ASEC earnings responses including too few low earners and too few very high earners. That is, compared to measures for the same individuals from administrative tax data, the Census hot-deck procedure based on the MAR assumption assigns missing earnings that are too high for low-

wage workers and too low for very high-wage workers, compressing the overall distribution of earnings. Earnings shares among the top 1% of earners are lower by at least 20 percent in the ASEC compared to matched administrative tax records, with about half accounted for by nonresponse and half to topcoding in the ASEC. This gap between the ASEC and data from the DER appears to have worsened in recent years, primarily among men. Our hybrid measures using ASEC earnings for CPS respondents and administrative DER earnings for nonrespondents produce intermediate estimates. Users of public ASEC data should incorporate both the new rank-swap topcode series recently made available by Census, and should incorporate the Pareto topcode adjustments to better account for upper-tail inequality.

I. Census Imputation and Topcoding Procedures and Implications for Inequality

Earnings (and income) nonresponse in the ASEC has increased over time, particularly following changes in the survey in 1994. Although this is widely known, less well known is that in addition to item nonresponse with respect to earnings and other CPS questions, there exists ASEC supplement nonresponse. This occurs when households participating and responding in the monthly CPS refuse to participate in the ASEC supplement. When this occurs, Census imputes the entire household record (so-called whole imputes) by replacing the blank supplement of the nonresponding household with the completed supplement of another household. Since the late 1990s, item nonresponse to the earnings question in the ASEC has been just over 20 percent and supplement whole imputes about 10 percent, resulting in total earnings nonresponse rates in excess of 30 percent (Bollinger et al. 2014; Hokayem et al. 2015).[1]

A. Imputations for Nonresponse

The Census Bureau has used a hot deck procedure for imputing missing income since 1962, with the current system in place with few changes since 1989 (Welniak 1990).[2] The ASEC uses a sequential hot deck procedure to address item nonresponse for missing earnings data by assigning individuals with missing earnings values that come from individuals ("donors") with similar characteristics. The ASEC sequential hot deck procedure for earnings variables first

[1] Another manifestation of earnings nonresponse is unit nonresponse, whereby the prospective sample member refuses or cannot be contacted for the initial CPS (monthly) survey. Dixon (2012) reports rates of unit nonresponse in CPS in the 8-9 percent range during our sample period. Korinek et al. (2007) provide evidence suggesting that there is negative selection into response, with households in higher income areas less likely to participate.

[2] The sequential hot deck procedures used in the March survey prior to 1989 were fairly primitive, with schooling not a match variable until 1975. Lillard, Smith, and Welch (1986) provided an influential critique of Census methods. Welniak (1990) documents changes over time in Census hot deck methods for the March CPS.

divides individuals with missing data into one of 12 allocation groups defined by the pattern of nonresponse (e.g., only missing earnings from longest job, or missing both longest job information and earnings). Second, an observation in each allocation group is matched to a donor with complete data based on a large set of socioeconomic match variables. If no match is found based on the large set of variables, then a match variable is dropped and variable definitions collapsed (i.e., categories are broadened) to be less restrictive. This process is repeated until a match is found. When a match is found, the missing earnings amount is replaced with the reported earnings from the first available matched donor.

The Census also uses a hot deck procedure for whole supplement nonresponse. Instead of 12 allocation groups, the whole imputation procedure uses 8 allocation groups. Moreover, the set of match variables is smaller than the set used for item nonresponse, consisting solely of variables from the basic monthly CPS. To be considered a donor for whole imputations, an ASEC respondent has to meet the minimum requirement that at least one person in the household has answered one of the following questions: worked at a job or business in the last year; received federal or state unemployment compensation in the last year; received supplemental unemployment benefits in the last year; received union unemployment or strike benefit in the last year; or lived in the same house one year ago. This requirement implies that whole supplement donors do not have to answer all the ASEC questions and can have item imputations. Similar to the sequential hot deck procedure for item nonresponse, the match process sequentially drops variables and makes them less restrictive until a donor is found.

B. Nonresponse and Inequality Measurement

Whether and to what extent unconditional (i.e. “raw”) or conditional (i.e. residual) measures of inequality using ASEC are affected by nonresponse depends on the type of nonresponse. Assuming that earnings nonrespondents are missing completely at random, raw inequality measures are expected to be unbiased and equivalent for large samples that include and exclude imputed earners. Under the MCAR assumption, the quality of the imputation procedure should not affect measures of unconditional inequality. Even with MCAR, however, the quality and specific details of the imputation procedure do affect measures of conditional or residual inequality owing to “match bias” (Hirsch and Schumacher 2004; Bollinger and Hirsch 2006). If covariates used in the hot deck imputation procedure are broader (more crude) than in the researcher’s model, measures of residual inequality will be larger using the full sample than

the sample that excludes imputed earners. For example, if the researcher conditions the earnings dispersion measure on worker industry and location (e.g., state), but these are not hot deck match criteria, inclusion of imputed earners causes residual inequality to be overstated.

Of course, the MCAR assumption does not hold. Nonresponse varies with respect to measurable demographic and geographic descriptors (e.g., race, city size), some of which are correlated with earnings. The more common and pertinent assumption for researchers and statistical agencies is MAR – missing at random conditional on measured covariates. Under MAR, unconditioned measures of inequality may differ between the full sample with imputations and a sample omitting imputed earners. And it is not clear a priori which of these two imperfect samples provides the better measure. The full sample is likely to provide a good measure of unconditional inequality if the covariates used in the imputation procedure provide an unbiased measure of earnings and maintain variance. Using only respondents (non-imputes) provides more accurate earnings responses, but risks bias (absent reweighting) to the extent that nonresponse rates differ across the earnings distribution, as we subsequently show. The full sample with imputes is not appropriate for examining conditional inequality, however, because the relationship between inequality and the multivariate correlations with respect to demographic, geographical, and job attributes not used (or used fully) in the imputation process will be biased (i.e., the "match bias" previously discussed). Retaining imputed earners does not account for nonresponse bias (NMAR) since individuals not reporting earnings are assigned earnings from donors who respond. Stated simply, nearly all imputation procedures assume that nonresponse is MAR with respect to covariates used in a given hot deck matching process. The exception would be a procedure that explicitly corrects for selection (nonresponse) bias.

The respondent-only sample has the important advantage of including only those for whom we observe earnings (we ignore other reporting and measurement issues), but has the disadvantage that it need not be a representative sample with respect to observables, some of which may be correlated with earnings and earnings dispersion. A straightforward way to remedy the non-representativeness of the respondent sample is to rebalance the sample using inverse probability weights (IPW), giving larger weight to those with attributes associated with high nonresponse (e.g., residing in a large urban area) and low weights to those with lower rates of nonresponse (white, high school graduates). The expectation under MAR is that a properly rebalanced respondent-only sample should provide measures of unconditional inequality similar

to that for a full sample including imputes. The rebalanced respondent sample, moreover, has the advantage of being appropriate for analysis of residual inequality, whereas a sample containing imputed earners is not, absent an imputation process that assigns earnings using the same set of covariates as in the researcher's analysis. For those using public use CPS files, rebalanced respondent-only samples from the CPS may be appropriate although not necessarily advantageous for analyses of unconditional inequality, while in most cases are essential for analyses measuring residual inequality under MAR.

The focus of our paper is to examine the more difficult issue of earnings not missing at random conditional on covariates (NMAR); that is, nonignorable nonresponse. If this is so, neither the full sample with imputes nor the respondent-only sample provides unbiased measures of earnings and earnings inequality. The full sample is flawed by some unknown degree of nonresponse bias, while at the same time subject to match bias in analyses of residual inequality. A rebalanced respondent sample is not subject to match bias but, like the full sample, may produce biased estimates of inequality (unconditional or conditional) due to unobserved differences in earnings among those who do and do not respond.

There is surprisingly limited knowledge of the pattern or strength of nonresponse bias in the CPS. Older studies using small CPS samples of married males matched to administrative earnings data (e.g., Greenlees et al. 1982 use a 1973 CPS sample) concluded that there was negative selection into nonresponse; that is, those with higher earnings, conditioned on covariates, being least likely to respond (Kline and Santos (2013) use this same 1973 CPS sample). Bollinger and Hirsch (2013), using a Heckman selection model with public use CPS files, also conclude that there is a central tendency toward negative selection (particularly so among men), but that this tendency is weak.

More recently, Hokayem, Bollinger, and Ziliak (2015) and Bollinger, Hirsch, Hokayem, and Ziliak (2014) examined nonresponse bias using internal ASEC files matched to administrative earnings data (DER), the data set used in this paper. In contrast to prior studies, they do not assume that there is common selection into nonresponse across the distribution. In their analysis, Hokayem et al. address the related issue of how nonresponse affects the measurement of poverty, finding that the official Census poverty rate is biased downward by about a percentage point due to nonresponse, meaning poverty is undercounted by about 3

million persons in a typical year. Bollinger et al. focus on residual nonresponse among full-time/full-year wage and salary workers. They conclude that nonresponse is U-shaped over the wage distribution, being constant over most of the distribution, but substantially higher in the left and right tails; that is, there exists strong positive selection into response in the left tail and strong negative selection in the far right tail. This evidence confirms the Lillard, Smith, and Welch (1986) conjecture of U-shaped nonresponse, for which Kline and Santos (2013) found support using data from the early 1970s. Bollinger et al., however, only examine full-time/full-year workers and do not examine how nonresponse affects the level and trends in inequality as we do here.

C. Measuring Inequality with Topcoding

Income and earnings variables in the CPS are topcoded not only in public use files, but also in internal Census files, though at substantially different levels. For example, the internal topcode for the person's earnings from longest job is $1,099,999, whereas in public-release versions of the ASEC it is $250,000. Prior to the 1996 survey year, Census only released the topcode value in the public versions of the data. Starting in 1996, they instead released the average value of earnings among those topcoded based on the broad groups of gender, race/ethnicity, and worker status.[3] The mean values were constructed based on earnings values between the public topcode value and the internal topcode value. Starting with the 2011 survey year, Census adopted a new approach called "rank proximity swapping," whereby they now order topcoded earners from lowest to highest and randomly swap out earnings between individuals within a bounded range (and again, below the internal topcode). Unlike the cell-mean series, this new approach preserves the distribution of earnings above the topcode.[4]

In our analysis we present several measures of inequality—the Gini coefficient, the top 1% share, and 90/50 and 50/10 percentile ratios. The Gini coefficient is a preferred summary measure of the entire distribution of earnings (Burkhauser et al. 2012), the share accruing to the top 1% is a focal measure in the recent administrative tax-record literature (Piketty and Saez 2003), and the percentile ratio has been a standard method in the residual inequality literature (Autor et al. 2008). The assignment of topcodes has a mechanical effect on summary inequality

[3] Larrimore et al. (2008) used internal ASEC data to construct the cell-mean series back to the 1976 survey year.
[4] Census has made available to the user community the rank-proximity swapped values for topcoded persons back to 1975 at https://www.census.gov/housing/extract_files/data%20extracts/income%20data%20files/. To further protect respondent confidentiality, Census rounds swapped values.

dispersion measures such as earnings shares and Gini – the higher the topcode value, the higher is measured inequality. By contrast, treatment of topcodes has no effect on percentile ratios as long as the topcoded value exceeds earnings at the percentile selected for the ratio's numerator.[5]

What can be disputed is the choice of an appropriate topcode adjustment. A typical approach by researchers using public use files is to apply a fixed topcode multiple for all workers and years, say multiplying the topcode by 1.4 or 1.5, which is intended to represent the mean level of earnings for those at and above the topcode based on the Pareto distribution. Somewhat problematic is the use of a single topcode multiple; at a minimum one might vary the topcode multiple based on differences by gender and year as we do below. Armour et al. (2014) recently provided a lucid discussion of problems with topcoded earnings in the CPS. As they highlight, the standard approach is to assume a two-point distribution to estimate the Pareto shape parameter, which we refer to as $\hat{\alpha}^{Baseline}$:

$$(1)\quad \hat{\alpha}^{Baseline} = \frac{\ln(\frac{C}{T})}{\ln(\frac{X_T}{X_C})}.$$

In equation (1) it is assumed that the earnings distribution is characterized by the Pareto above a lower cutoff value, X_C, and a second, and higher cutoff value, X_T, is required to estimate the shape parameter. Choice of these two points is somewhat arbitrary, but an obvious one for the former is the earnings of the 99th percentile, and for the latter is the earnings value of the topcode value. In this case, *C* then represents the number of persons with earnings above the lower cutoff (i.e. the 99th percentile), and *T* is the number of persons with earnings above the higher cutoff (i.e. the topcoded value). Armour et al. (2014) suggest a more robust maximum likelihood estimator of the shape parameter ($\hat{\alpha}^{MLE}$) that utilizes the actual earnings of individuals between the two cut points (in fact the sum of (log) earnings between the points), and not just the two points themselves, defined as

$$(2)\quad \hat{\alpha}^{MLE} = \frac{M}{\{T\ln(X_T)+\sum_{X_M\le x_i<X_T}\ln(x_i)+(M+T)\ln(X_C)\}},$$

[5] Likewise, bottom codes also have a predictable effect; for example researcher choices regarding such things as the lowest wage measures to include (say, none below the minimum wage), the inclusion of zero earnings, and the inclusion of earnings losses among the self-employed.

where M is the number of persons with earnings between the 99th percentile and the topcode, $\ln(x_i)$ is the natural log of earnings of individual i, and $\ln(X_C)$ and $\ln(X_T)$ are the same as in equation (1).

Like Armour et al., our internal ASEC data set allows us to observe topcodes higher than in the public use files. In addition, however, we also have non-topcoded measures of earnings from linked administrative DER records. In our analysis, we construct the alternative measures of inequality using the DER, the ASEC earnings reports with and without imputed values, and ASEC data with three different adjustments to the internal topcode—a fixed multiple of 1.4, and the Pareto shape parameters in equations (1) and (2). The Pareto shape parameters are estimated separately by year and sample, both using the internal ASEC alone and the ASEC in conjunction with the DER.

II. Data Description

The data used in our analysis are ASEC person records matched to the DER file for survey years 1998-2011, reporting earnings for calendar years 1997-2010. Our estimation sample includes all wage and salary and self-employed workers, ages 18-64, not enrolled in school. We separately provide analyses using the full sample, men alone, women alone, and only full-time, full-year (FT/FY) workers (men and women combined). We identify FT/FY workers based on annual hours worked, the product of weeks worked (WKSWORK) and usual hours worked per week (HRSWK), requiring that a FT/FY worker has worked at least 50 weeks in the prior year and at least 35 hours per week. All estimates are weighted by the ASEC supplement weight.

A. Detailed Earnings Records and ASEC-DER Match Rates across the Distribution

The DER file is an extract of SSA's Master Earnings File and includes data on total earnings, including wages and salaries subject to taxation under the Federal Insurance Contributions Act (FICA) and earnings from self-employment subject to taxation under the Self-Employment Contributions Act (SECA). Only positive self-employment earnings are reported in DER because individuals do not make SECA contributions if they have self-employment losses (Nicholas and Wiseman 2009). The DER file contains all earnings reported on workers' W-2 tax forms (and Form 1099 if self employed). These earnings are not capped at the FICA contribution amounts and include earnings not covered by Old Age Survivor's Disability Insurance but subject to the Medicare tax. Unlike ASEC earnings records, the DER earnings are not topcoded.

This is important given that there are substantial concerns regarding nonresponse and nonresponse bias in the right tail of the distribution, but knowledge on these issues is quite limited. The DER file also contains deferred wages such as contributions to 401(k), 403(b), 408(k), 457(b), and 501(c) retirement plans, as well as health savings accounts. However, as described in Abowd and Stinson (2013), some forms gross compensation are not in DER such as pre-tax health insurance premiums and education benefits. Of potentially greater concern, particularly for the left tail of the earnings distribution, is that DER cannot measure earnings that are off-the-books, or are not subject to FICA taxation such as public employees in certain states with solely state-funded pension plans, and thus not reported to SSA.

Workers in the DER file are identified by a Protected Identification Key (PIK), a confidentiality-protected version of the Social Security Number (SSN) assigned by Census. DER files are matched to ASEC files by the Census Bureau's Center for Administrative Records Research and Applications (CARRA). Since the CPS no longer asks respondents for a SSN, CARRA uses its own record linkage software system, the Person Validation System, to assign a model-based SSN (Wagner and Layne 2014). This assignment relies on a probabilistic matching model based on name, address, date of birth, and gender. The SSN is then converted to a PIK. The SSN from the DER file received from SSA is also converted to a PIK, and the two files are matched based on the PIK and do not contain SSN.

[Figures 1a and 1b here]

Figure 1 shows how the DER match rates vary across the ASEC earnings distribution, with Figure 1a depicting match rates in the first half of the sample period 1997-2004 and Figure 1b for the second half spanning 2005-2010. We split the sample in presenting the match rates because Census changed its consent protocol for matching respondents to administrative data beginning with the 2006 ASEC. Prior to this CPS collected respondent Social Security Numbers and an affirmative agreement allowing a match to administrative data; i.e., an "opt-in" consent option, with "no consent" being the default. Beginning with survey year 2006, this changed to a consent default coupled with an "opt-out" option – respondents not wanting to be matched to administrative data had to notify the Census Bureau through the website or use a special mail-in response. If the Census Bureau doesn't receive this notification, the respondent is assigned a SSN using the Person Validation System. Comparing Figures 1a and 1b, the switch to a consent

default increased the match rate by 20 percentage points in most of the distribution, from 60-70 percent to 80-90 percent. However, regardless of default or the sample (gender and FT/FY), the match patterns across the distribution are similar—match rates are lower for those in the left tail of the distribution, but rates vary little throughout the rest of the distribution.

B. Earnings Nonresponse across the DER Distribution

Workers can appear in the DER files more than once each year if they have several jobs. We collapse the DER records into one earnings observation per worker per year by aggregating total earnings from Box 1 of the W-2 (labeled "Wages, tips, other compensation") across each worker's wage and salary employers, plus the higher of Box 3 (Social Security earnings) and Box 5 (Medicare wages and tips) across each worker's self-employment earnings sources. We also collapse total deferred contributions. That is, earnings and total contributions are summed across all of an individual's employers and businesses. The wage and salary portion of total DER earnings is most compatible with CPS earnings from all wage and salary jobs (ERN-VAL plus WSAL-VAL). The self-employed portion of total DER earnings most closely corresponds to SE-VAL and the farm portion to FRM-VAL. We classify a worker as having imputed ASEC earnings (i.e., nonresponse) if any component of wages and salary, self-employed, or farm earnings is imputed, or the entire ASEC supplement is imputed (flag FL-665).

[Table 1 here]

Table 1 provides descriptive statistics for the full sample, for both men and women separately, and for the narrower sample of FT/FY workers over the entire 1997-2010 period. Earnings are reported in constant 2012 dollars using the Personal Consumption Expenditure deflator. For the full sample, just under 25% of workers have earnings imputed (item plus whole nonresponse), more than 5 percentage points less than the ASEC nonresponse rate not conditioned on the DER match, indicating that those not matched to the DER exhibit higher ASEC nonresponse. Roughly 6% of workers have topcoded earnings in the internal ASEC (from any earnings source), but this is far more prevalent among men than women (10% versus 2%). Mean ASEC and DER earnings are shown separately for workers who are ASEC respondents and nonrespondents. The ASEC earnings shown for nonrespondents include the Census imputed values. In contrast, DER earnings shown for ASEC nonrespondents (and respondents) are individuals' administrative earnings and not imputed values. In all cases, mean earnings for

respondents and nonrespondents, using both ASEC and DER and for men and women, are reasonably close in value. The one exception is among men, where DER earnings of nonrespondents exceeds that of respondents by $2,000.

[Figure 2 here]

ASEC nonresponse may be higher than average among those difficult to match based on information provided to Census and IRS, or workers with earnings off-the-books. Such individuals are likely to be concentrated in the tails of the DER earnings distribution. Indeed this appears to be the case, as seen in Figure 2, which depicts the unconditional nonresponse rate for each of the four samples across each group's DER earnings distribution.[6] There we see that the nonresponse rate (item and whole) is U-shaped. Nonresponse is substantially higher in the lower and extreme upper tails of the distribution. High rates of right-tail nonresponse are limited to earnings above the CPS public-use topcode values. The U-shaped nonresponse pattern implies that nonresponse bias is not a constant, but varies in magnitude and form across the distribution. In the lower tail of the wage distribution there is positive selection into survey response, those with unusually low earnings being most likely not to report earnings. This is most pronounced among men and FT/FY workers. In the far right tail of the distribution there is negative selection into response, those with very high earnings being most likely to not report earnings. We note that this pattern is unchanged if we examine the sub-periods of 1997-2000, 2001-2004, and 2005-2010 (not shown in the figure), suggesting that the U-shaped nonresponse is not an artifact of business-cycle contractions nor of the "opt-in" versus "opt-out" ASEC-DER merge process. Moreover, Bollinger et al. (2014) find a similar U-pattern of nonresponse, conditioned on covariates, in a sample of FT/FY wage and salary workers. Although nonignorable nonresponse (NMAR) does not appear to be a serious issue over most of the distribution, the finding that there exists "trouble in the tails" leaves open the possibility that nonresponse has substantive effects on measures of inequality. Indeed, although not shown in Table 1, the standard deviation of earnings for nonrespondents is substantially higher than for respondents, consistent with nonrespondents being overrepresented (underrepresented) in the tails (middle) of the earnings distribution. We turn to such evidence below.

[6] Because each group has a different distribution, earnings values at each percentile differ by group. Percentile earnings values are higher for men than women, and higher for the FT/FY workers than for the full sample.

III. Evidence on Nonresponse and Topcoding on Levels and Trends in Earnings Inequality

We begin our analysis of the effects of nonresponse and topcoding on earnings inequality by presenting trends in the Gini coefficient for four earnings measures: (1) the ASEC with Census imputes included; (2) the ASEC with Census imputes excluded; (3) the DER for all matched respondents and nonrespondents; (4) and the ASEC for respondents and DER for matched nonrespondents. In constructing the latter two DER measures we use ASEC earnings (including imputes) for those workers without a DER match in order to keep the sample composition the same across measures. All inequality series are unconditioned, i.e. do not control for covariates. We present results first for the full sample of workers, and then examine heterogeneity in patterns among the three subsamples.

[Figure 3 here]

In Figure 3 we show the earnings Gini for the full sample of workers. Shown in diamonds with a blue line is the full ASEC sample, in squares with an orange line is the ASEC with imputes excluded, in triangles with a gray line is the DER for all matched workers (ASEC for non-matched), and in crosses with a yellow line is the ASEC for responders and DER for matched nonrespondents (ASEC for all others). Comparing the full ASEC with imputes (diamonds) versus respondents only (squares), one sees that the respondent-only sample shows too low a level of inequality owing to the omission of nonrespondents disproportionately represented in the far left and right tails. Hence, omission of imputes is inappropriate for measuring unconditioned inequality, absent a proper reweighting of the respondent sample that gives heavy weight to low and very high earners as we discuss below.[7]

As compared to the two DER measures, the ASEC measures show different levels and trends in earnings inequality. Earnings inequality in the ASEC is largely flat over the sample period, and everywhere below the DER. Using DER earnings for the sample of workers, we find a higher degree of inequality and an upward trend over the period. Note that some of the upward trend coincides with the change in the CPS-DER matching procedure in 2005 from a default opt-out to default opt-in. The increase in inequality in the middle of the decade, however, also coincides with trends in general IRS Form 1040 tax data (Saez 2015), suggesting that the DER

[7] Recall that if the focus were on residual inequality, it might not be appropriate to include imputed earners since correlation with covariates omitted from the hot deck match would be attenuated (Bollinger and Hirsch 2006).

trend is not primarily due to the broader DER sample resulting from the new opt-in procedure. The hybrid DER measure that uses DER earnings for nonrespondents and ASEC earnings for respondents produces a Gini level roughly halfway between the pure ASEC (with Census imputations) and DER measures. Comparing the hybrid measures to the pure ASEC measures supports the conclusion that response bias (NMAR) causes an understatement in the level and trend in earnings inequality based solely on ASEC.

[Figure 4 here]

We next turn to trends in the share of earnings accruing to the top 1% of workers in Figure 4, the most prominent inequality measure presented from tax data (Piketty and Saez 2003). Here we see pronounced differences between the ASEC and the DER. Among all workers, there is a modest downward trend in the top 1% share in the ASEC, and a slight upward trend in the DER. Averaged over all years, the DER measure of the top centile is 2.1 percentage points higher than the ASEC measure, or 21% higher than the ASEC mean share of 9.95%. This gap grew over time, with the DER-ASEC gap averaging 1.7 percentage points in the first half of the sample period, and 2.7 percentage points in the second half. There are two key takeaways from Figures 3 and 4. First, ASEC measures of inequality tend to understate inequality because the Census hot deck (owing to nonresponse bias) imputes earnings for nonrespondents that are too high in the left tail and too low in the right tail, thus understating inequality. Second, even with DER earnings assigned to nonrespondents, the Gini based solely on DER values (triangles) is systematically higher than the hybrid series (crosses). Thus, there is greater variability shown in DER than in ASEC reported earnings.

[Figures 5-6 here]

In Figures 5 and 6 we present 50/10 and 90/50 percentile ratios, both as a way to isolate upper-half from lower-half inequality and to mitigate undue influence of very high or very low earnings that may be present in the Gini and top 1% shares (Autor et al. 2008). The percentile ratios avoid use of earnings in the most extreme tails where there exist the highest rates of nonresponse and the most noisy reporting of earnings, particularly so in the left tail where some earnings are off the books or not subject to FICA taxation and many have worked minimal hours during the prior year. In the right tail, the 90^{th} percentile is well below the topcodes in both the

ASEC internal and public CPS files. The downside is that the percentile ratios do not reflect just how low and how high are earnings in the tails.

The 50/10 trends in Figure 5 show clear-cut declines in the late 1990s across all four earnings series, reflecting more rapid growth in earnings at the 10th percentile than the median. After 2000, the two ASEC series see-sawed up, then down, and then up again with the onset of the Great Recession in 2007. The DER series, however, maintained a similar but more steady pattern after 2000. While earnings at both the 10th and 50th percentiles declined during the Great Recession, the decline was more pronounced at the 10th percentile, leading to an uptick in lower-tail inequality in recent years (though still lower than at the start of the sample period).

The 90/50 ratio for all workers in Figure 6, on the other hand, shows an overall increase in upper-half inequality over the sample period across all four measures. Focusing on the late 1990s, the declining 50/10 coupled with the flat or increasing 90/50 ratios reflected continuing hollowing out of occupations and earnings in the middle of the distribution. There are no doubt multiple explanations, among them technological change, increased globalization, and the decline in private sector unionization. Polarization in the labor market has been shown to be associated with task-based skilled biased technological change due to information technology, which in turn results in employment and earnings declines in many middle class occupations with programmable (routinizable) tasks, while having lesser effects on low-skill service occupations that involve tasks not readily programmable (Autor, Levy, and Murnane 2003). That said, polarization slowed markedly after about 2000 (Autor 2015), with increasing inequality since that time driven by earnings growth at the top (such a conclusion also follows based on a combination of Figures 5 and 6). This increase is most pronounced in the DER series, and as with the Gini coefficients in Figure 3, the hybrid DER series lies in between the full ASEC and the DER, suggesting that nonresponse accounts for a substantive share of the difference in inequality across survey and administrative data.

A. Alternative Earnings Measures using the DER and ASEC

We next consider several refinements on our earnings measures to explore further the role of nonresponse and topcoded earnings in the ASEC. We focus initially on the DER, where we present three alternatives to the prior series. First, in the prior graphs we replaced the ASEC with the DER for any worker with a match to the DER regardless of imputation status, or

alternatively, for matched nonrespondents. Because the latter two groups include a convolution of nonrespondents and topcoded workers, it is less obvious what direct role the topcode in the internal ASEC plays vis-à-vis administrative tax data. To examine this, we compare the baselone DER series to a series where the DER is used only for topcoded ASEC values with a DER match (in both cases the ASEC is used when a DER match is not available to hold sample composition constant). Second, because the DER does not capture earnings off-the-book (or for a small number of workers not subject to FICA/SECA taxation), the higher level of inequality observed in the DER might be an artifact of underreported earnings in the lower half of the distribution. To test this, we replace the ASEC with the DER for workers in the top but not bottom half of the ASEC earnings distribution, regardless of imputation or topcode status. Third, we use the DER to extend the hot deck procedure to predict ASEC earnings for nonrespondents and those topcoded. Specifically, we run the following regression

$$(3)\quad y_{it}^{ASEC} = \alpha_t + \beta_t y_{it}^{DER} + X_{it}\gamma_t + u_{it},$$

where y_{it}^{ASEC} refers to earnings in the ASEC, y_{it}^{DER} is earnings in the DER, and X_{it} is a vector of demographics including a quartic in age, and indicators for race, education, industry, and occupation. We estimate this model for matched respondents who are not topcoded, and then replace the imputed or topcoded ASEC values with the fitted value $\hat{y}_{it}^{ASEC}$ for matched nonrespondents and matched topcoded workers (and use the actual ASEC reports for all others). In effect, this approach extends the MAR assumption to the case where ASEC earnings are missing (or topcoded) at random conditional on covariates *and* the DER.

[Figures 7-8 here]

In Figures 7 and 8 we present the Gini coefficients and top 1% earnings shares, respectively, for the all-DER earnings and the three alternatives using information from the DER to supplement ASEC values. A common theme in Figures 7 and 8 is that topcoded earnings alone in the internal ASEC are not the primary cause of the gap in inequality estimates from tax data in the DER versus ASEC survey data. The DER-only series (diamonds) shows substantially higher and rising inequality as compared to ASEC earnings with DER replacing ASEC topcodes (shown in squares). In addition, the majority of the gap between the DER and ASEC earnings inequality arises from earnings in the upper half of the ASEC distribution, and not from off-the-books underreporting in the lower half. This conclusion is based on the minimal differences

between the DER-only series (diamonds) and the hybrid ASEC-DER series with DER earnings replacing the ASEC in the top half of the ASEC distribution (triangles). Moreover, our use of DER earnings in a regression-based ASEC hot deck does not ameliorate violations of MAR, producing estimates of inequality much lower than those seen directly using DER earnings (compare DER earnings in diamonds to the Predicted ASEC(DER) series in crosses).

We next return to the internal ASEC to examine possible "fixes" to the topcode. Specifically, we consider three variants of the Pareto shape parameter—a fixed multiple of 1.4, the time-varying baseline Pareto estimate shown previously in equation (1), and the time-varying MLE Pareto estimate in equation (2). In the case of the fixed multiple, if any of the four earnings components, i.e. ERN-VAL, WS-VAL, SE-VAL, and FRM-VAL, are topcoded then we replace the topcode with 1.4 times that value, and aggregate the adjusted series to construct total earnings. For the other two approaches to Pareto shape parameters, we only estimate the shape of the distribution for ERN-VAL because the number of topcoded individuals in the other three earnings components are too few to provide reliable estimates. As a consequence, for the baseline and MLE Pareto series we apply the fixed multiple of 1.4 to topcoded values in WS-VAL, SE-VAL, and FRM-VAL, but use the time-varying shape parameters estimated in equations (1) and (2) for ERN-VAL. As before, we aggregate across the adjusted components to arrive at a revised total earnings to construct the inequality measures.

[Table 2 here]

Table 2 presents the estimated baseline and MLE shape parameters using the internal ASEC for the full sample and the three subsamples. There it is clear that the fixed multiple of 1.4 provides a substantial underestimate of the shape of the distribution in the upper tail in nearly all years and samples. Moreover, with few exceptions, the MLE version provides a larger estimate than the baseline. In the full sample (bottom line), the average baseline Pareto parameter is 1.76 and the MLE is 1.99; in the subsamples they are 1.75(1.91), 1.77(1.94), 1.79(2.43) for the full-time/full-year, men, and women samples, respectively (with the MLE average in parentheses).[8]

[8] We also considered two additional variants to estimating the Pareto parameters, each using a combination of the ASEC with the DER. In one case, we replaced the estimated 99th percentile (X_C) from the ASEC in equation (1) and (2) with the corresponding 99th percentile estimated from the DER. In the second case, we also replaced the fixed topcode value from the internal CPS with a value derived from the DER. Specifically, for each year and sample, we computed the fraction of persons topcoded in ERN-VAL in the ASEC, and then computed the corresponding (1-x)% percentile from the DER distribution. For example, if the sample has 10,000 persons, and 10 are topcoded, then

[Figures 9 and 10 here]

Figures 9 and 10 present estimates of the Gini coefficients and the top 1% shares using the three Pareto shape parameter adjustments. For sake of comparison, we also present the corresponding estimates from the full ASEC and the DER regardless of imputation status. A different story emerges in the two figures. In Figure 9, the Pareto adjustments do little to close the gap between the ASEC and the DER in terms of summary inequality measures like the Gini. In Figure 10, however, it appears that the Pareto adjustments, especially the MLE variant from equation (2), provide a considerable improvement over the ASEC in estimating upper-tail inequality. Indeed, in several years the Pareto MLE adjustment provides estimates of the top 1% share comparable to those in DER (and in most years does better than a series that replaces ASEC with DER earnings for nonrespondents, but not respondents). This suggests that researchers focusing on upper-tail inequality should incorporate the Pareto shape parameters in Table 2 into their analyses. This approach could be readily incorporated into the official Census estimates of inequality and released as a separate series on an annual basis.

B. Subsample Analyses

In this section we examine heterogeneity in the base-case full-sample estimates of the Gini and top 1% shares in Figures 3 and 4 by restricting our sample to FT/FY workers, men, and women in Figures 11a-11c and 12a-12c, respectively. Among these subsamples, we find the same relative rankings and same movements over time as seen previously for all workers in Figures 3 and 4; that is, the DER provides the largest estimates of inequality, the ASEC with nonrespondents dropped provide the lowest, and the DER for nonrespondents only (and ASEC for all others) splits the difference between the full ASEC and the full DER. There are some notable differences, however. First, the level of inequality is systematically lower for the FT/FY sample for whom hours worked over the year varies far less than among all workers. Second, when we examine the Gini solely among ASEC respondents (the squares), we obtain the lowest level of inequality among the alternative series and subgroups. This is not surprising given our showing in Figure 2 that FT/FY nonrespondents disproportionately come from the tails of the distribution. Third, without exception, the level of inequality among men exceeds that among

0.1% of the sample is topcoded. We then estimated the 99.9 percentile in the DER and used that as the topcode value X_T in equations (1) and (2). In both cases, the estimated Pareto shape parameters were little changed, suggesting that the internal ASEC provides robust estimates of the Pareto shape of the administrative tax data.

women, though the trends are similar across the earnings series. This is particularly notable in the top 1% shares in Figures 12b and 12c. Among men the average gap in the top 1% share between the ASEC and the DER is 3 percentage points for the whole period (or 30% over the ASEC mean of 105), and sizable 5.2 percentage points alone in 2010. Fourth, there is little difference between the full ASEC among women, and the ASEC without nonrespondents, which again is consistent with the relatively constant nonresponse rates among women compared to men. Based on our analysis, we conclude that it is important to include nonrespondents in analyses of unconditional inequality (but not necessarily residual inequality, given concerns about match bias and evidence in Lemieux 2010), and that Census imputation methods fail to fully account for the true dispersion of inequality among the nonrespondents due to response bias in the tails.

[Figures 11a-12c here]

IV. "Fixes" for the Public ASEC

The general CPS user community does not have access to either the internal ASEC used in this paper (and by Census employees), or the DER. The advantage of the former comes primarily from data with higher topcode values compared to the public ASEC. Since 1996 Census has attempted to address this discrepancy, while still maintaining confidentiality, by releasing proxy values for those individuals with earnings in between the public and internal topcodes. During survey years 1996-2010 the proxy came in the form of cell means, while from 2011 onward via rank swapping. Recently Census released rank-swap values for all the topcode income components (not just earnings) back to 1975. In Figure 13 we demonstrate that in the case of earnings, rank swapping matters for inequality estimates, especially upper-tail inequality.

[Figure 13 here]

Figure 13 depicts the estimated top percentile in the public ASEC for the full sample using the version released by Census denoted by "Cell Mean" and a version that replaces the cell mean with the rank swap value (denoted by "Rank Swap" in the figure). We make this replacement for each of the individual earnings components—ERN-VAL, WS-VAL, SE-VAL, and FRM-VAL—and then aggregate up to person-level earnings. Figure 13 shows the inflation-adjusted top percentile varies wildly across years using the cell-mean approach, compared to the relatively stable rank-swap estimate. The implication is that inequality measures such as the top

1% share vary considerably in public ASEC data under the cell mean approach compared to rank swap. For example, in 2000 the rank swap estimate of the top 1% share is 9.9%, whereas it is only 1.9% in the cell-mean series. The effect on summary measures like the Gini, however, is negligible. The conclusion we draw is that researchers using ASEC data prior to the 2011 survey year should incorporate the rank swap series into their data, especially for research that can be heavily influenced by data in the right tail of the distribution.

We next compare the internal ASEC inequality estimates to several alternatives in the public ASEC. These alternatives include the rank-swap public ASEC data, the rank-swap public ASEC with the Pareto MLE estimates from Table 2 incorporated for those topcoded observations (note that the topcodes in the rank swap series are the same as the internal ASEC), the rank-swap public ASEC with nonrespondents dropped but with the sample reweighted with inverse probability weights (IPW Public ASEC), and the rank-swap public ASEC with IPW and Pareto MLE topcodes.[9] Figure 14 presents the Gini coefficients, and Figure 15 the top 1% earnings shares.

[Figures 14-15 here]

Both figures show that starting in 2007 we get the same estimates of inequality with the rank-swap public ASEC as with the internal ASEC, though in years prior we estimate slightly lower inequality in the public data. The figures also show that with rank-swap topcodes, along with Pareto MLE adjustments to the topcode value, summary inequality measures and upper-tail measures are higher in the public data than the internal data. This suggests that users of public data should incorporate these Pareto MLE topcode adjustments. Finally, inverse probability weighted (rank swap) public ASEC data on respondents results in inequality estimates—both Gini and top 1% shares—that are quite comparable to the full public sample with nonrespondents included. This is consistent with our conjecture that properly reweighted respondent-only sample should be similar to the sample with imputes included under the MAR assumption, and furthermore suggests that for analyses of unconditional inequality researchers should not drop

[9] For the IPW model we estimate a flexible probit of the probability of response as a function of a quartic in age, gender, race, marital status, education, employment type (e.g. federal, private), self-employed status, full-time employment status, hours and weeks of work, nativity, occupation, who in the household responded to the survey, metro size, region of country, interactions of full-time status with the other demographics, and year dummies. We then multiply the ASEC supplement weight with the inverse predicted probability of response.

nonrespondents but for analyses of residual inequality the IPW measure is preferred.[10]

V. Conclusion

Measures of U.S. earnings (and income) inequality from both statistical agencies and researchers rely heavily on the CPS-ASEC. Yet a substantial and increasing share of individuals and households surveyed in the CPS either do not participate in the ASEC supplement or participate but refuse to report earnings, instead having their earnings imputed by Census. These imputations rely on nonresponse being missing at random (MAR); that is, nonresponse being uncorrelated with true earnings, conditioned on measured covariates.

In this paper, we used ASEC data for calendar years 1997 to 2010 matched to administrative tax records in order to explore the nature of nonresponse bias and how such bias affects measures of the level and trends in earnings inequality. We find strong evidence that there exists little nonresponse bias over most of the earnings distribution, but bias does exist in the tails. Nonresponse is found to be U-shaped with respect to administrative earnings (observed for nonrespondents and respondents), being flat over most of the distribution, but with high nonresponse in roughly the lowest 10 percentiles of the distribution and in the very highest percentiles. In the left (far right) tail, those with unusually low (high) earnings are least likely to respond. This relationship holds both conditional and unconditioned on covariates.

The effect of such nonresponse bias is that earnings inequality is understated, with ASEC earnings including too few low earners and too few very high earners. Earnings imputations for nonrespondents do not correct (or totally correct) this bias since imputed values are based on the earnings of responding donors with similar measured (but not unmeasured) attributes. Hence, imputed values are too high among nonrespondents in the left tail and too low for nonrespondents in the far right tail.

We confirm that earnings inequality measures based solely on ASEC are systematically lower than are measures based solely on administrative earnings. Hybrid measures using ASEC earnings for CPS respondents and administrative DER earnings for nonrespondents produce estimates roughly midway between the other two measures. Although inequality trends over the 1997-2010 period are not linear and somewhat noisy, we tend to find a slight upward trend in

[10] This is consistent with the approach by Armour et al. (2014) and Burkhauser et al. (2012), who have not omitted nonrespondents in their work using internal ASEC data.

inequality using the administrative earnings measure versus a slight downward trend relying solely on ASEC earnings.

References

Abowd, John M. and Martha H. Stinson. 2013. "Estimating Measurement Error in Annual Job Earnings: A Comparison of Survey and Administrative Data," *Review of Economics and Statistics*, 95(5): 1451-1467.

Armour, Philip, Richard V. Burkhauser, and Jeff Larrimore. 2014. "Using the Pareto Distribution to Improve Estimates of Topcoded Earnings," NBER Working Paper 19846.

Autor, David H. 2015. "Why Are There Still So May Jobs? The History and Future of Workplace Automation," *Journal of Economic Perspectives*, 29(3): 3-30.

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

Autor, David H., Frank Levy, and Richard J. Murnane. 2003. "The Skill Content of Recent Technological Change: An Empirical Exploration," *Quarterly Journal of Economics*, 118(4): 1279-1333.

Bollinger, Christopher R. and Barry T. Hirsch. 2006. "Match Bias from Earnings Imputation in the Current Population Survey: The Case of Imperfect Matching," *Journal of Labor Economics*, 24(3): 483-519.

Bollinger, Christopher R. and Barry T. Hirsch. 2013. "Is Earnings Nonresponse Ignorable?" *Review of Economics and Statistics*, 95(2): 407-416.

Bollinger, Christopher R., Barry T. Hirsch, Charles Hokayem, and James P. Ziliak. 2014. "Trouble in the Tails: Earnings Nonresponse and Response Bias across the Distribution," presented at Joint Statistical Meetings, Boston, August.

Bond, Brittany, J. David Brown, Adela Luque, and Amy O'Hara. 2013. "The Nature of the Bias When Studying Only Linkable Person Records: Evidence from the American Community Survey," Proceedings of the 2013 Federal Committee on Statistical Methodology (FCSM) Research Conference.

Bound, John, and George Johnson. 1992. "Changes in the Structure of Wages in the 1980s: An Evaluation of Alternative Explanations," *American Economic Review*, 82(3): 371-392.

Burkhauser, Richard V., Shuaizhang Feng, Stephen Jenkins and Jeff Larrimore. 2012. "Recent Trends in Top Income Shares in the USA: Reconciling Estimates from March CPS and IRS Tax Return Data," *Review of Economics and Statistics*, 94(2): 371-388.

Card, David, and John E. DiNardo. 2002." Skill-Biased Technological Change and Rising Wage Inequality: Some Problems and Puzzles," *Journal of Labor Economics*, 20(4): 733-783.

Chetty, Raj, Nathaniel Hendren, Patrick Kline, and Emmanuel Saez. 2014, "Where is the Land of Opportunity? The Geography of Intergenerational Mobility in the United States," *Quarterly Journal of Economics*, 129(4): 1553-1623.

DiNardo, John, Nicole Fortin, and Thomas Lemieux. 1996. "Labor Market Institutions, and the Distribution of Wages, 1973-1992: A Semi-parametric Approach," *Econometrica*, 64(5): 1001-1044.

Dixon, John. 2012. "Using Contact History Information to Adjust for Nonresponse in the Current Population Survey." In *JSM Proceedings*, Section on Government Statistics. Alexandria, VA: American Statistical Association, 1977-1982.

Greenlees, John, William Reece, and Kimberly Zieschang. 1982. "Imputation of Missing Values when the Probability of Response Depends on the Variable Being Imputed," *Journal of the American Statistical Association*, 77(378): 251-261.

Heckman, James J., and Paul A. LaFontaine. 2006. "Bias-Corrected Estimates of GED Returns," *Journal of Labor Economics*, 24 (3): 661-700.

Hirsch, Barry T., and Edward J. Schumacher. 2004. "Match Bias in Wage Gap Estimates Due to Earnings Imputation," *Journal of Labor Economics*, 22(3): 689-722.

Hokayem, Charles, Christopher R. Bollinger, and James P. Ziliak. 2015. "The Role of CPS Nonresponse in the Measurement of Poverty," *Journal of the American Statistical Association*, 110(511): 935-945.

Juhn, Chinhui, Kevin M. Murphy, and Brooks Pierce. 1993. "Wage Inequality and the Rise in the Returns to Skill," *Journal of Political Economy*, 101(3): 410-433.

Kline, Patrick, and Andres Santos. 2013. "Sensitivity to Missing Data Assumptions: Theory and an Evaluation of the U.S. Wage Structure," *Quantitative Economics*, 4 (2): 231-267.

Kopczuk, Wojciech, Emmanuel Saez, and Jae Song. 2010. "Earnings Inequality and Mobility in the United States: Evidence from Social Security Data Since 1937," *Quarterly Journal of Economics*, 125(1): 91-128.

Korinek, Anton, Johan A. Mistiaen, and Martin Ravallion. 2007. "An Econometric Method of Correcting for Unit Nonresponse Bias in Surveys," *Journal of Econometrics*, 136 (1): 213-235.

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.

Lemieux, Thomas. 2006. "Increasing Residual Wage Inequality: Composition Effects, Noisy Data, or Rising Demand for Skill," *American Economic Review*, 96(3): 461-498.

Lemieux, Thomas. 2010. “What Do We Really Know About Changes in Wage Inequality?” In *Labor in the New Economy*, K. Abraham, J. Spletzer, and M. Harper, (editors), University of Chicago Press, 17-59.

Lillard, Lee, James P. Smith, and Finis Welch. 1986. “What Do We Really Know about Wages? The Importance of Nonreporting and Census Imputation,” *Journal of Political Economy*, 94 (3): 489-506.

Little, Roderick J.A. and Donald B. Rubin. 2002. *Statistical Analysis with Missing Data, Second Edition*. Wiley-Interscience: Hoboken, NJ.

Nicholas, Joyce and Michael Wiseman. 2009. “Elderly Poverty and Supplemental Security Income,” *Social Security Bulletin*, 69(1): 45-73.

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

Saez, Emmanuel. 2015. “Striking it Richer: The Evolution of Top Incomes in the United States (Updated with 2013 preliminary estimates,” University of California, Berkeley, Mimeo.

Spletzer, James. 2014. “Inequality Statistics from the LEHD,” Presented at the Federal Economic Statistics Advisory Committee. Available at http://www.census.gov/fesac/pdf/2014-06-13/Spletzer_Background.pdf

Roemer, Mark. 2002. “Using Administrative Earnings Records to Assess Wage Data Quality in the Current Population Survey and the Survey of Income and Program Participation.” Longitudinal Employer-Household Dynamics Program Technical Paper No. TP-2002-22, U.S. Census Bureau.

Rubin, Donald B. 1976. “Inference and Missing Data,” *Biometrika*, 63(3): 581-592.

Wagner, Deborah, and Mary Layne. 2014. “The Person Identification Validation System (PVS): Applying the Center for Administrative Records Research and Applications’ (CARRA) Record Linkage Software,” CARRA Working Paper Series #2014-01, U.S. Census Bureau.

Welniak, Edward J. 1990. “Effects of the March Current Population Survey's New Processing System On Estimates of Income and Poverty,” Proceedings of the American Statistical Association.

Figure 1a. DER Match Rate Across the ASEC Earnings Distribution, 1997-2004

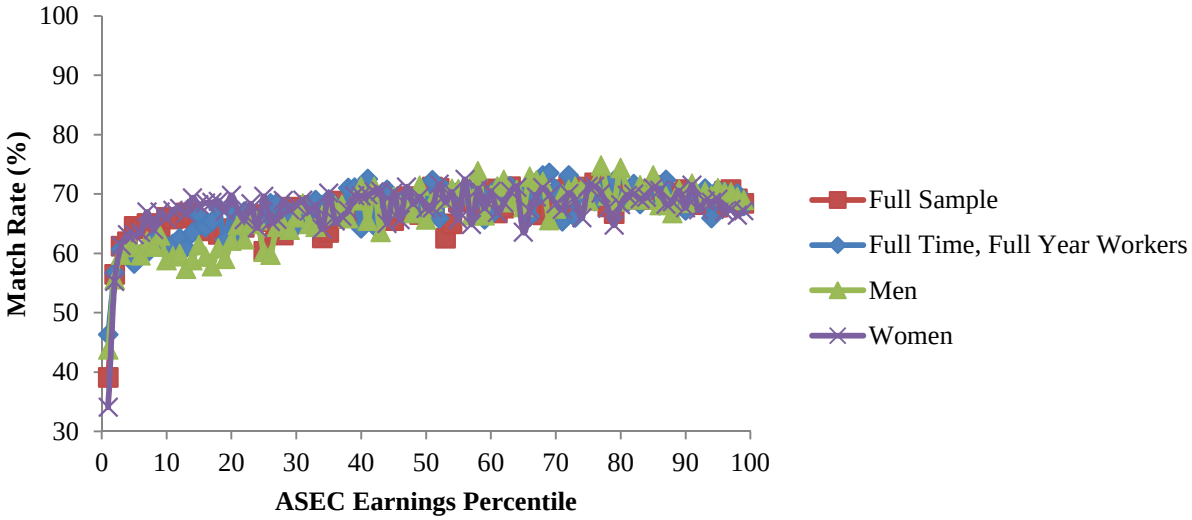

Figure 1b. DER Match Rate Across the ASEC Earnings Distribution, 2005-2010

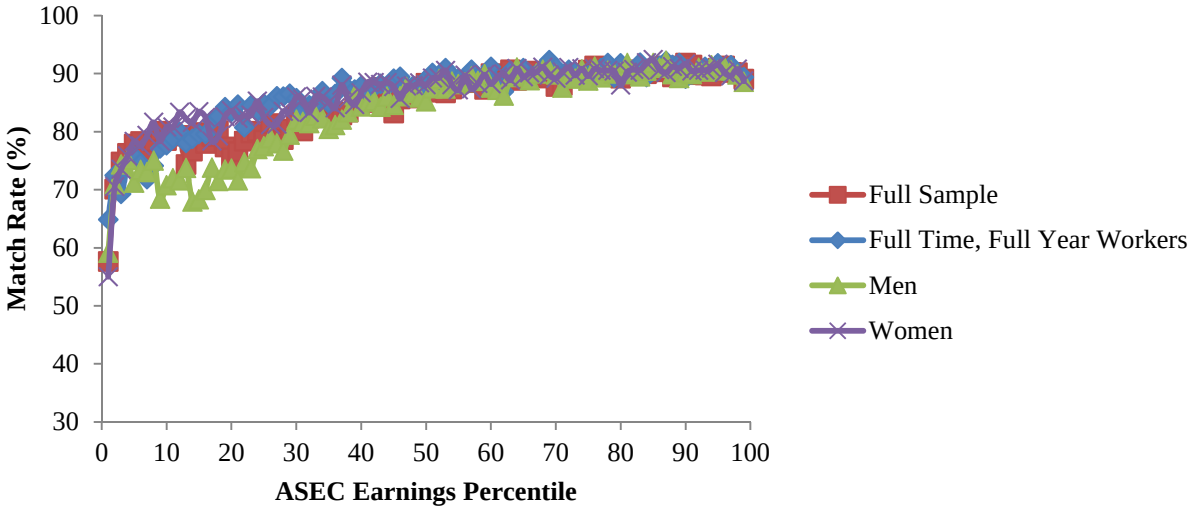

Figure 2. Item and Whole ASEC Earnings Nonresponse Rate Across the DER Earnings Distribution, 1997-2010

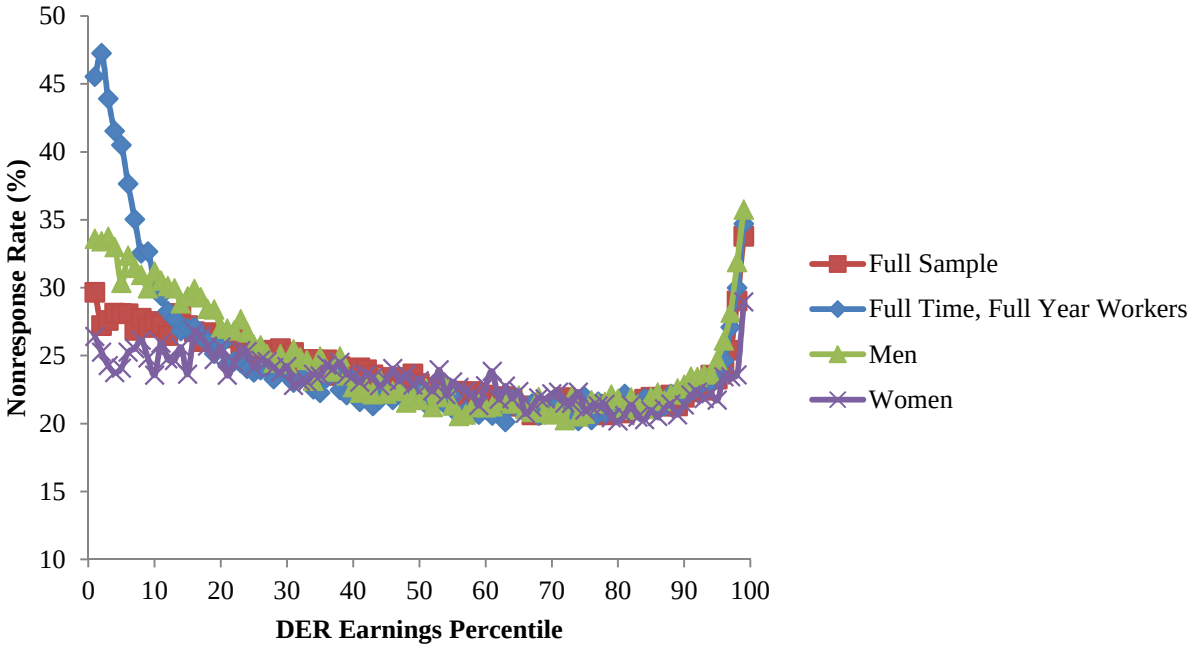

Figure 3. Trends in Gini Earnings Inequality, All Workers

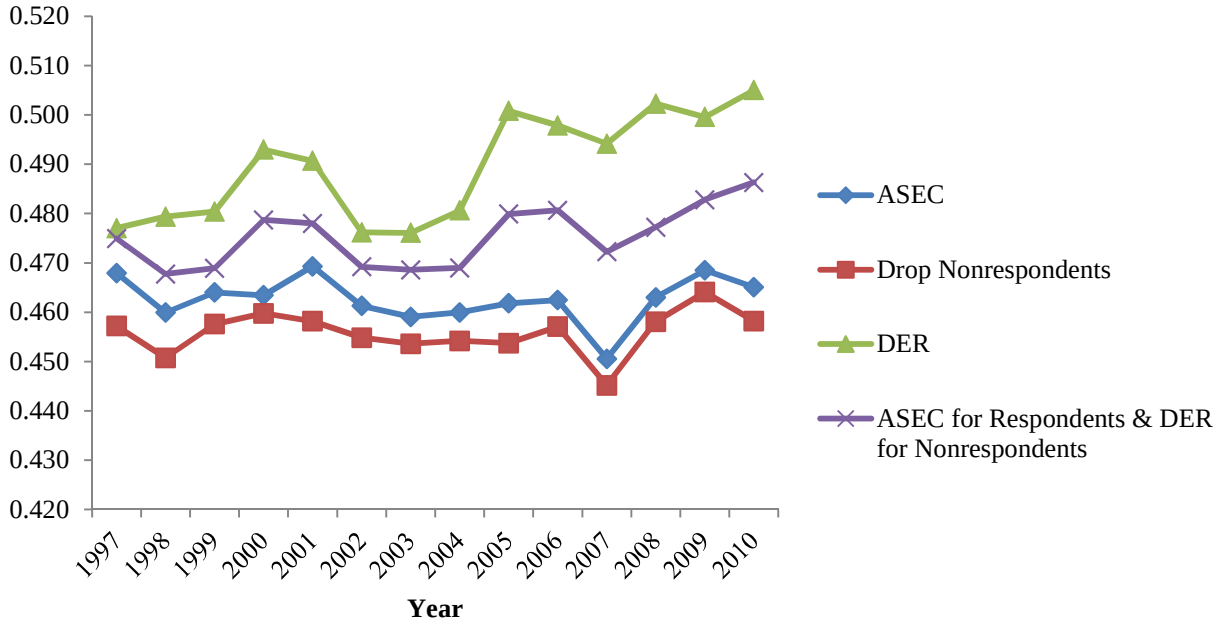

Figure 4. Trends in Top 1% Earnings Share, All Workers

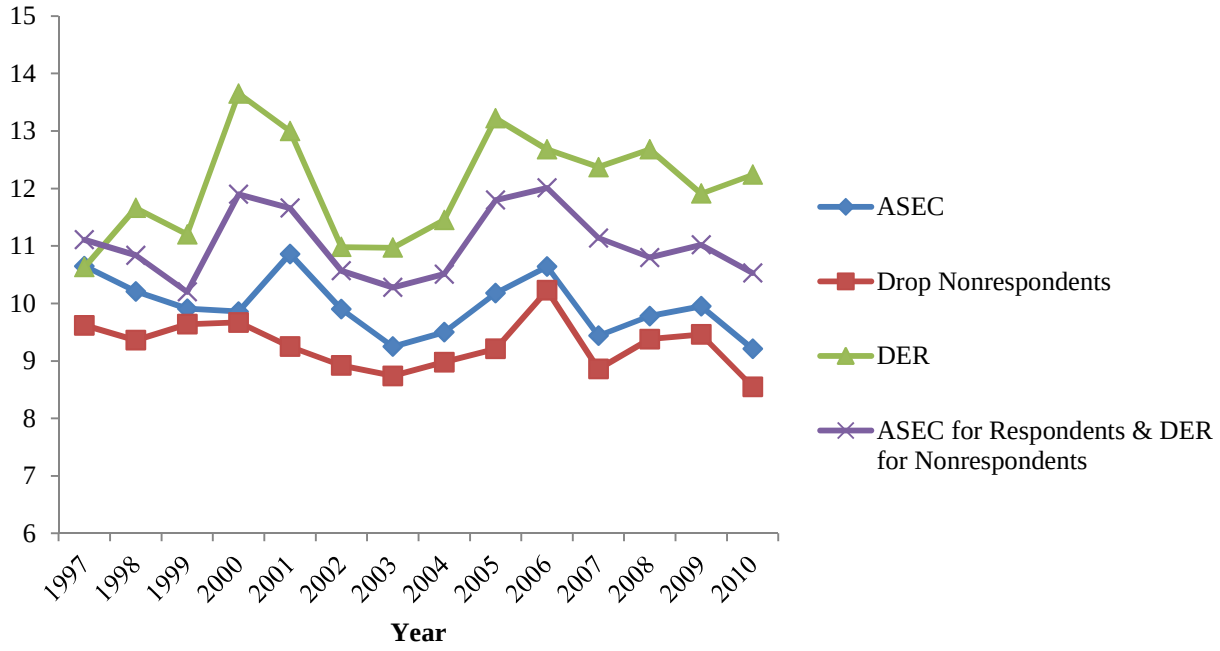

Figure 5. Trends in 50/10 Earnings Inequality, All Workers

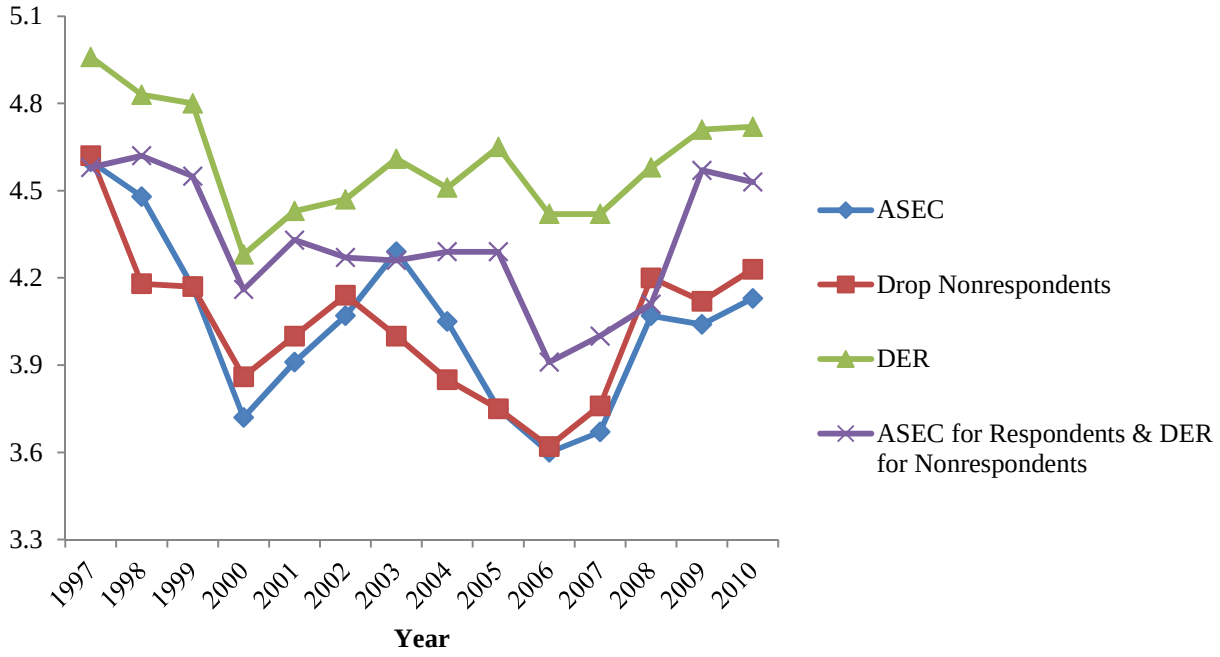

Figure 6. Trends in 90/50 Earnings Inequality, All Workers

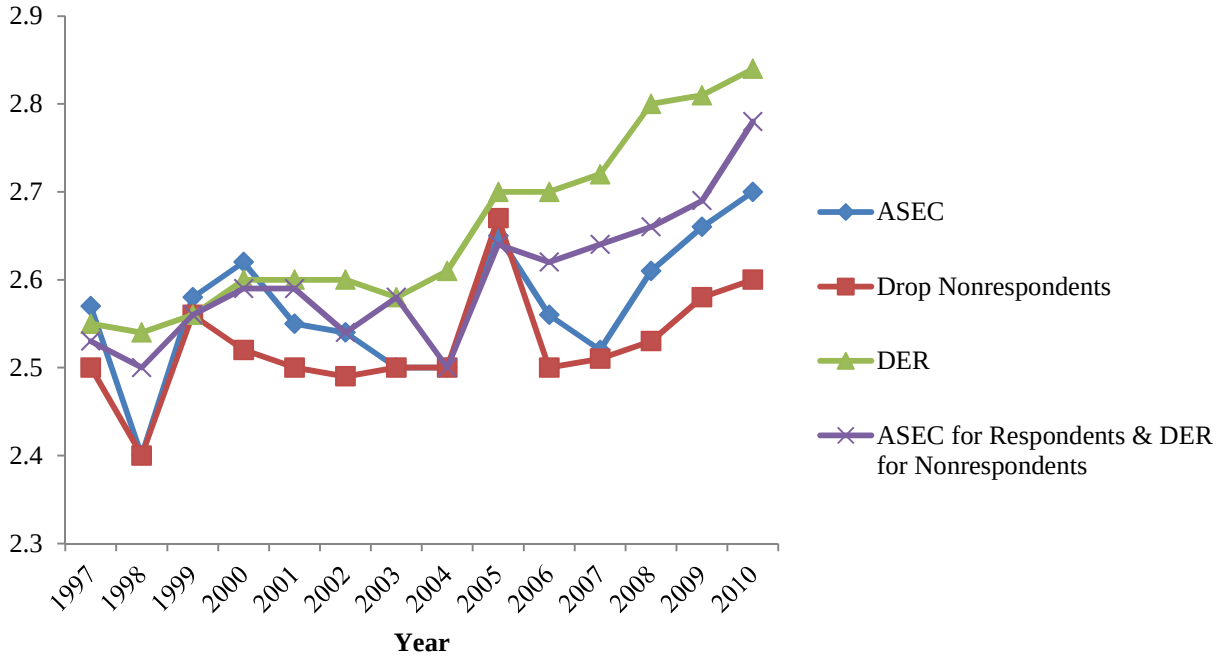

Figure 7. Trends in Gini Earnings Inequality for Alternative Hybrid ASEC-DER Measures, All Workers

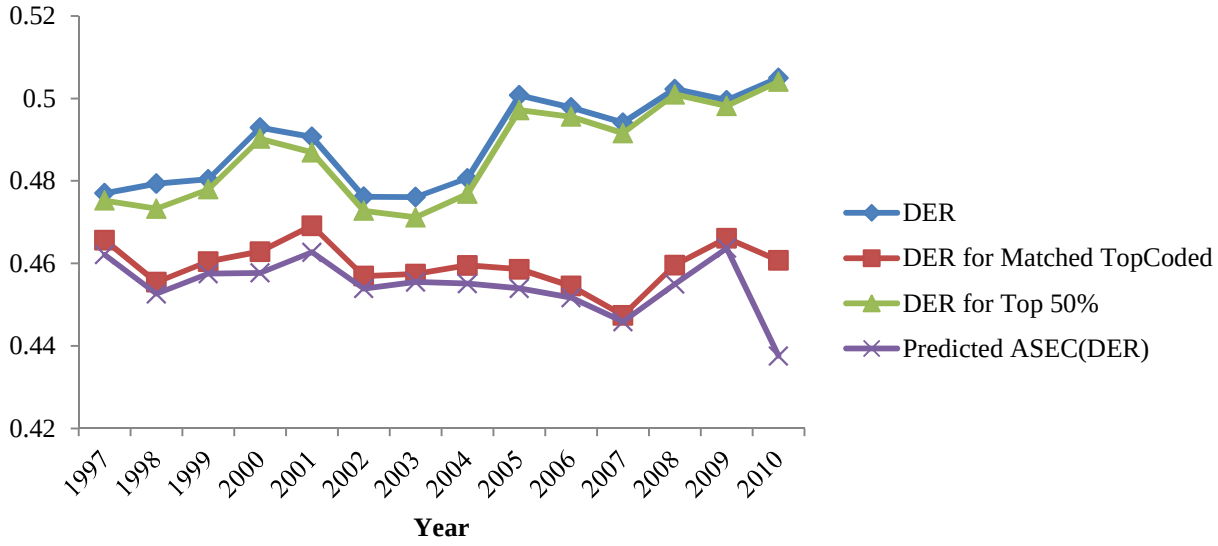

Figure 8. Trends in Top 1% Earnings Share for Alternative Hybrid ASEC-DER Measures, All Workers

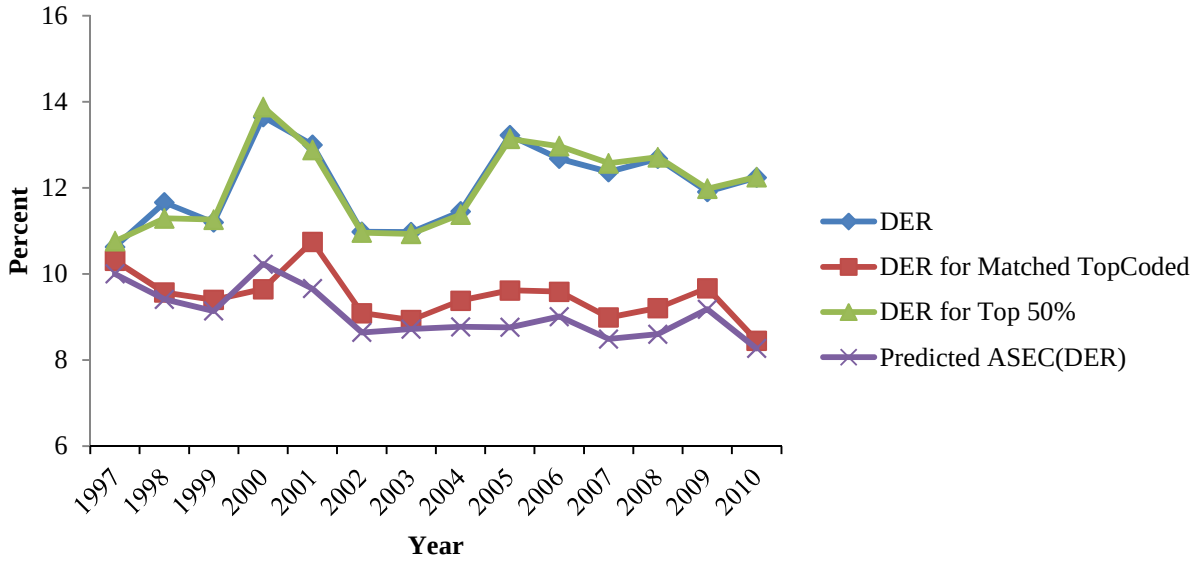

Figure 9. Trends in Gini Earnings Inequality for Alternative ASEC Topcode Measures, All Workers

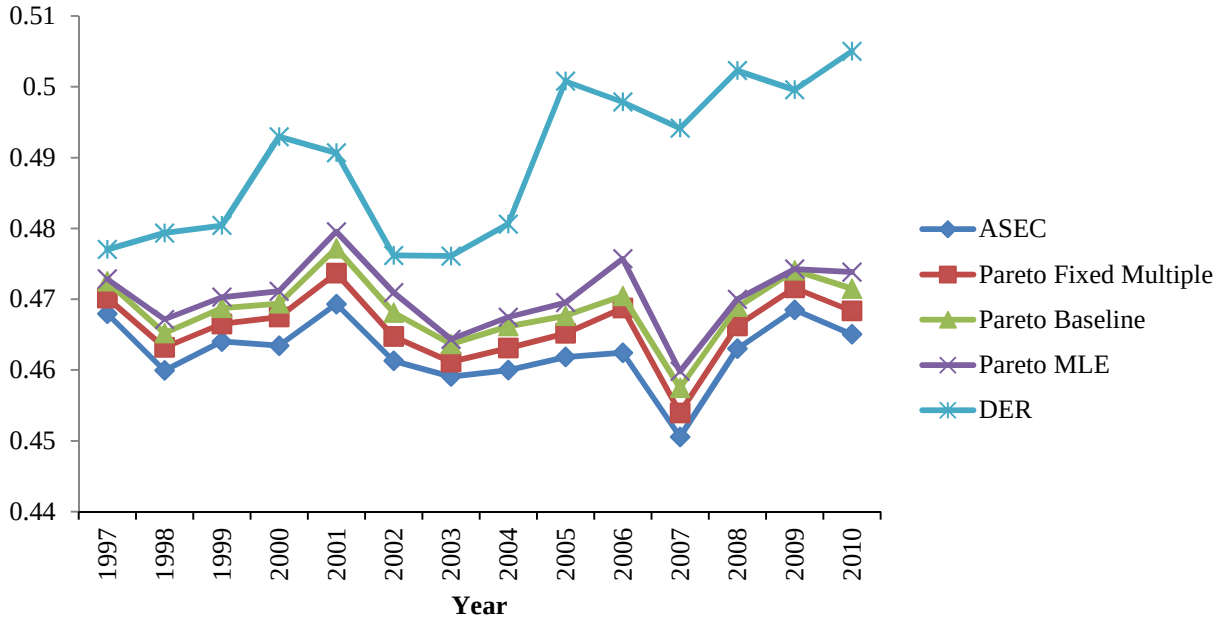

Figure 10. Trends in Top 1% Earnings Share for Alternative ASEC Topcode Measures, All Workers

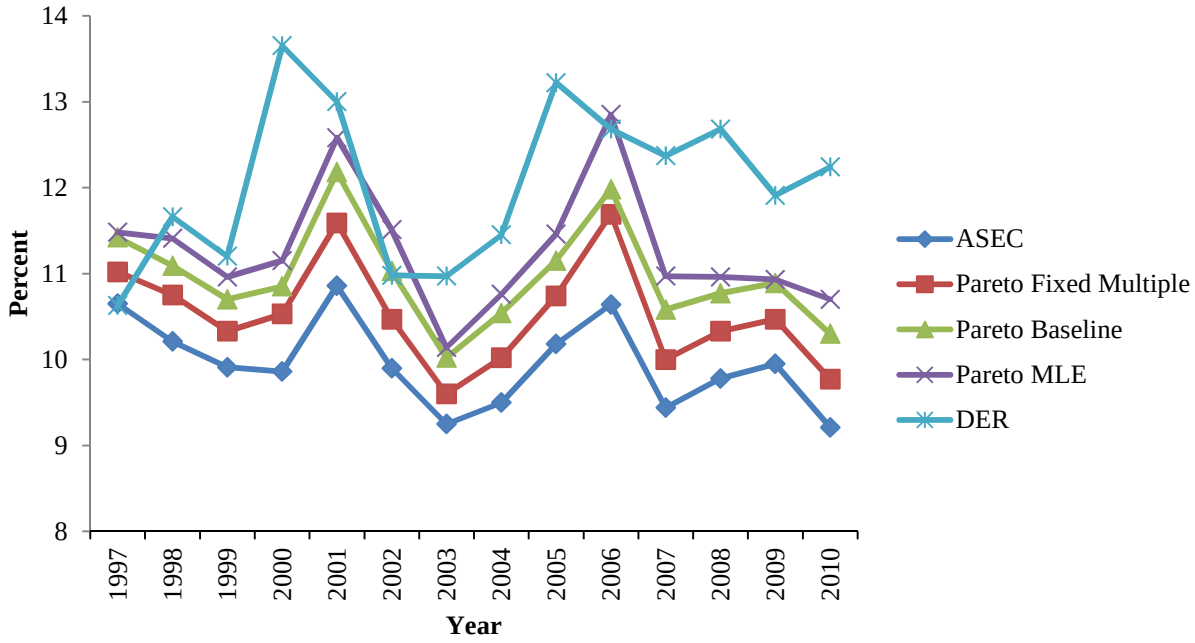

Figure 11a. Trends in Gini Earnings Inequality, Full-Time Year-Round Workers

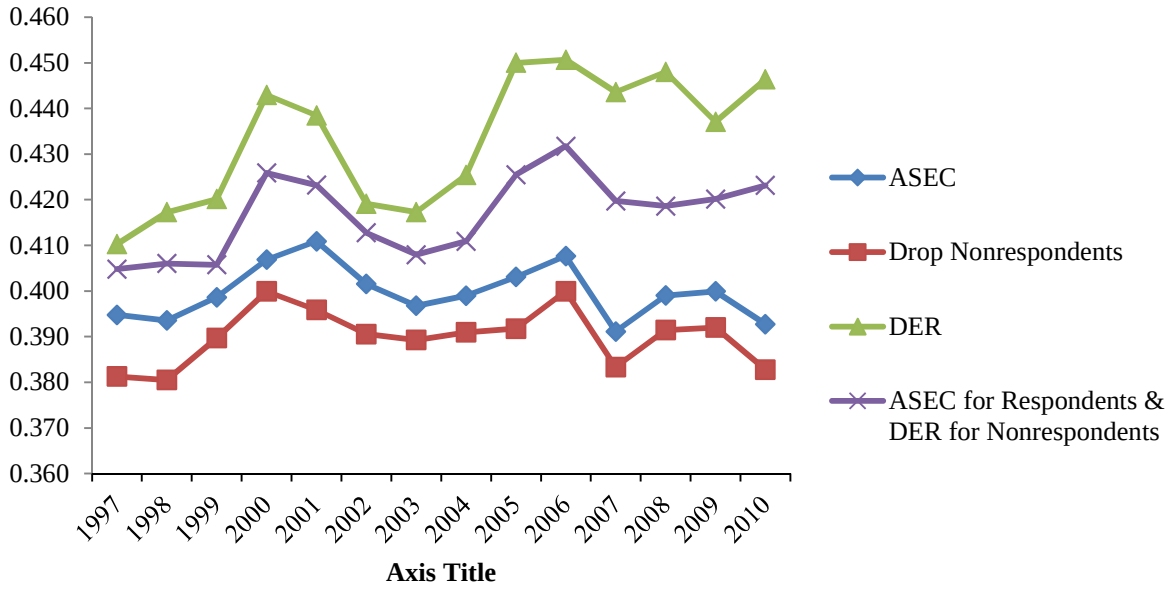

Figure 11b. Trends in Gini Earnings Inequality, Male Workers

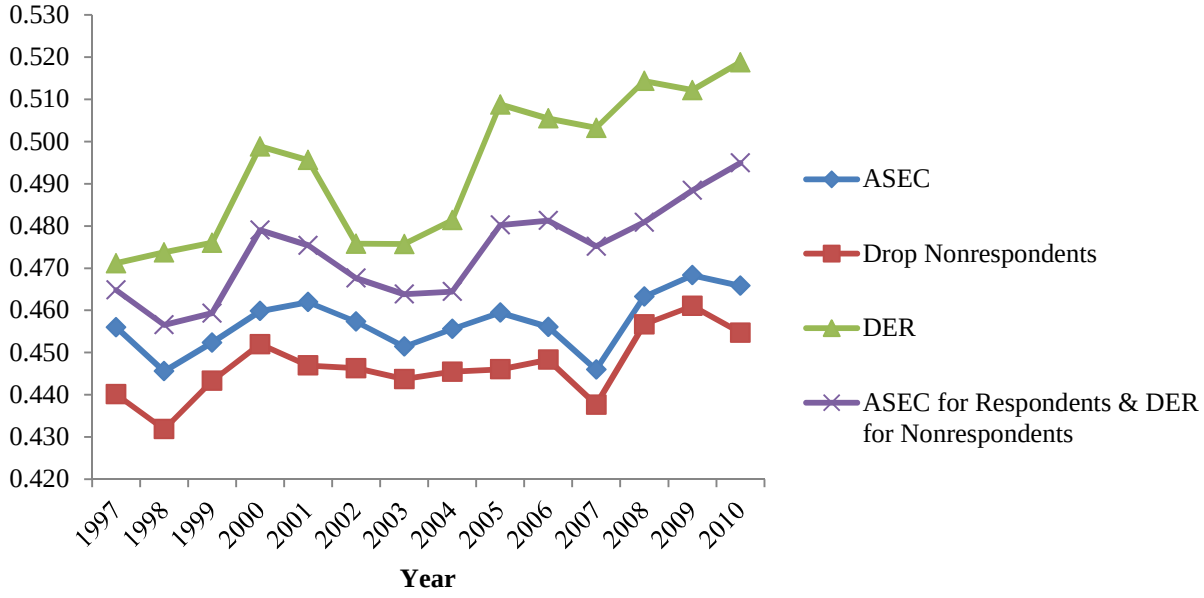

Figure 11c. Trends in Gini Earnings Inequality, Female Workers

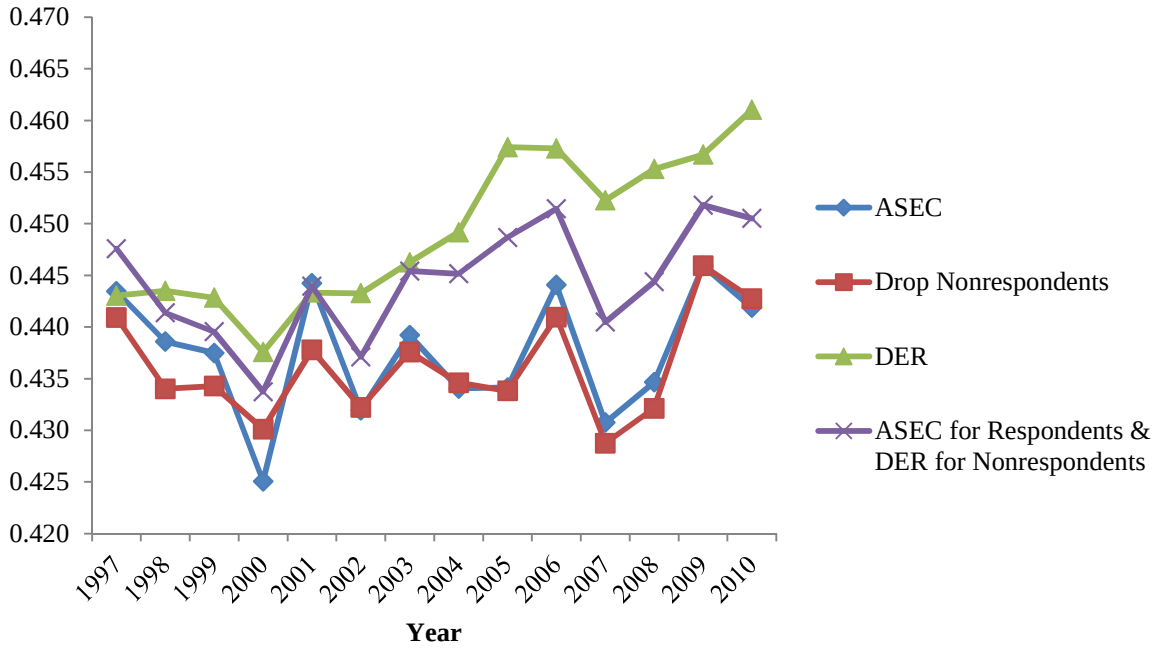

Figure 12a. Trends in Top 1% Earnings Share, Full-Time Full-Year Workers

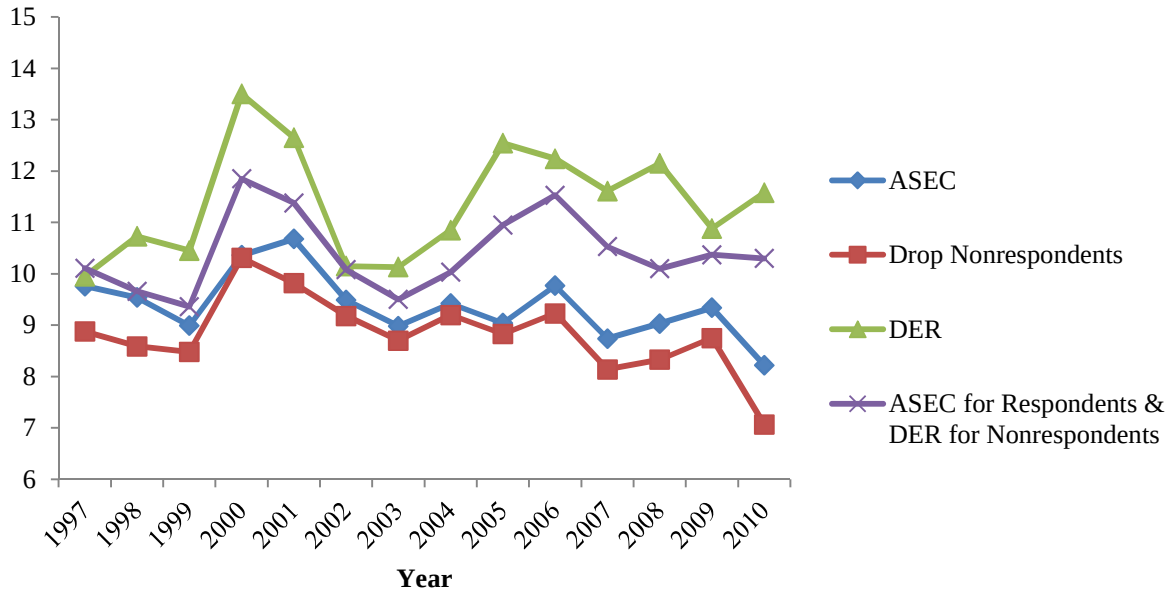

Figure 12b. Trends in Top 1% Earnings Share, Male Workers

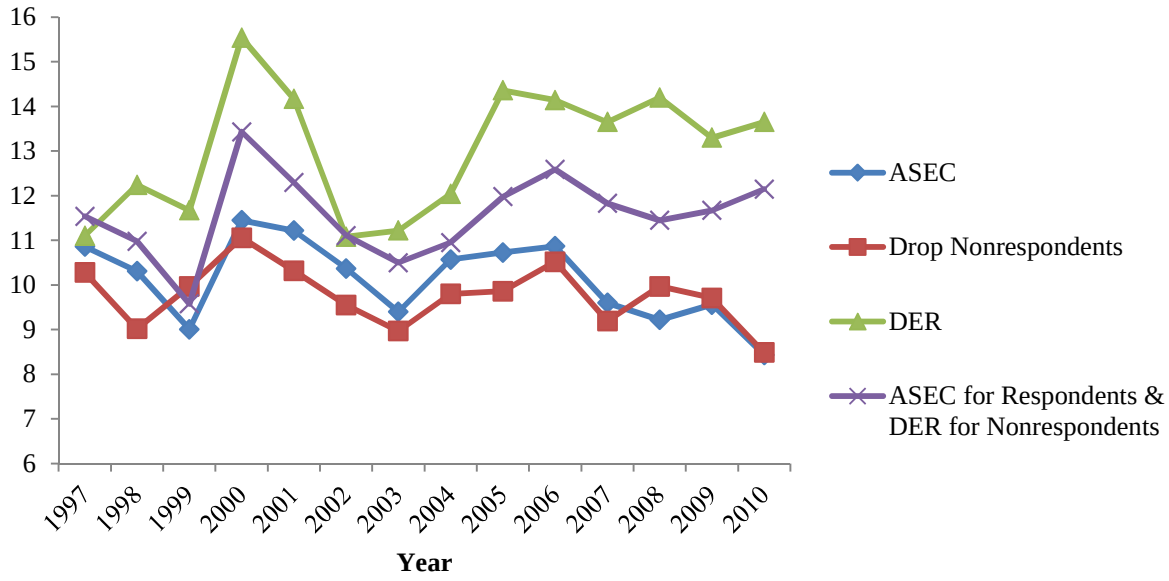

Figure 12c. Trends in Top 1% Earnings Share, Female Workers

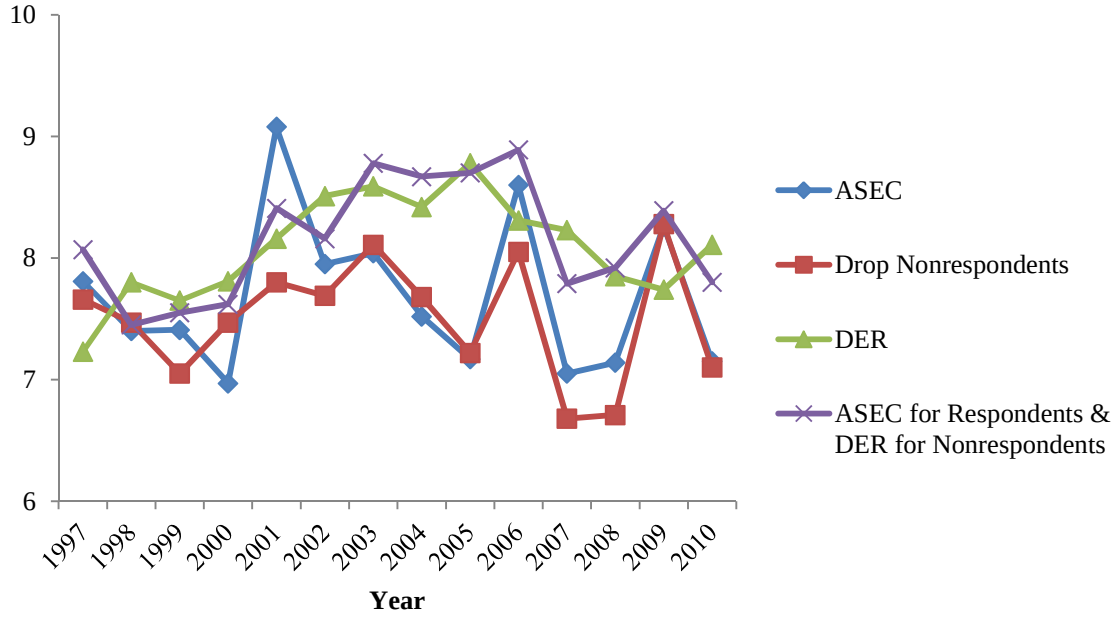

Figure 13. Trends in 99th Percentile in Public ASEC using Cell Mean and Rank Swap Approach to Topcoded Earnings

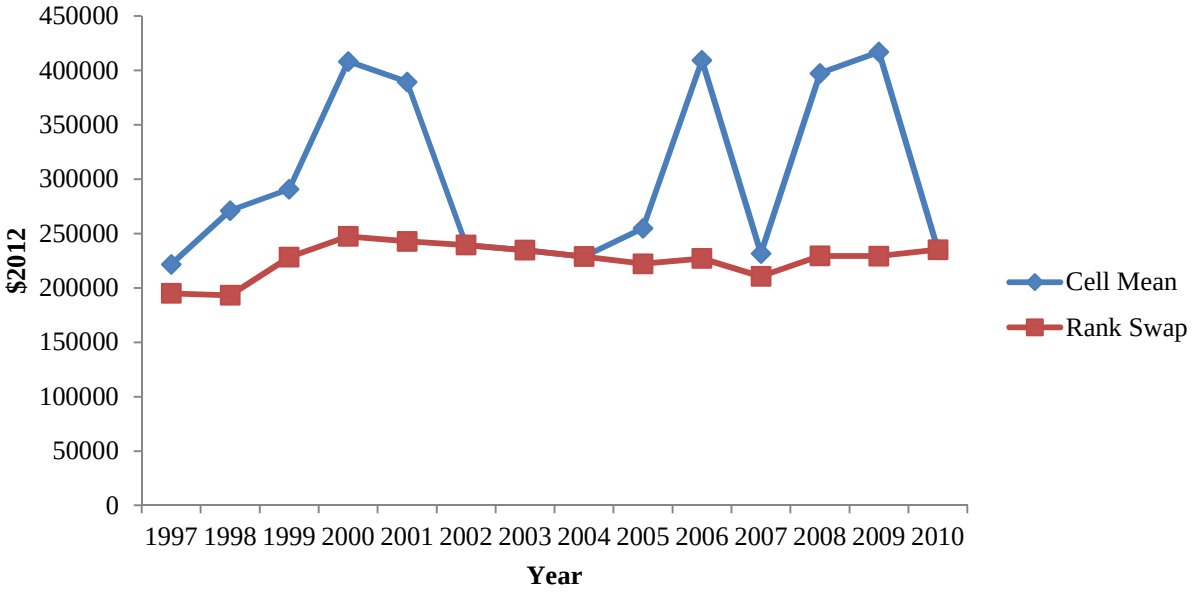

Figure 14. Trends in Gini Earnings Inequality in Public ASEC, All Workers

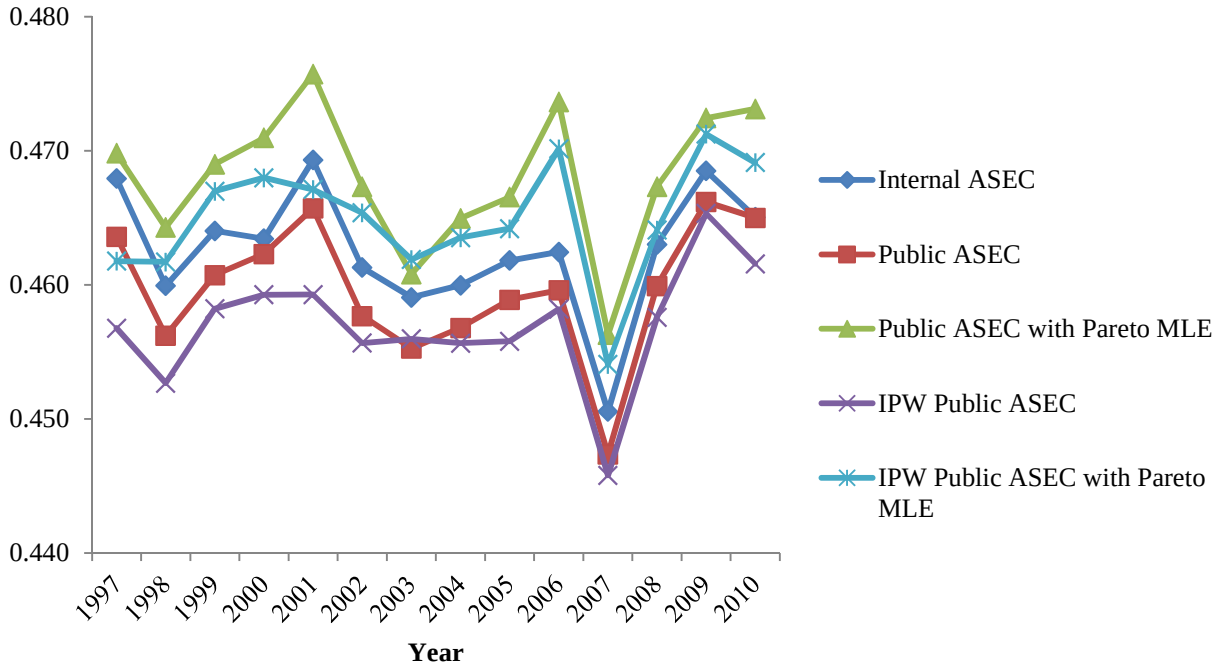

Figure 15. Trends in Top 1% Earnings Share in Public ASEC, All Workers

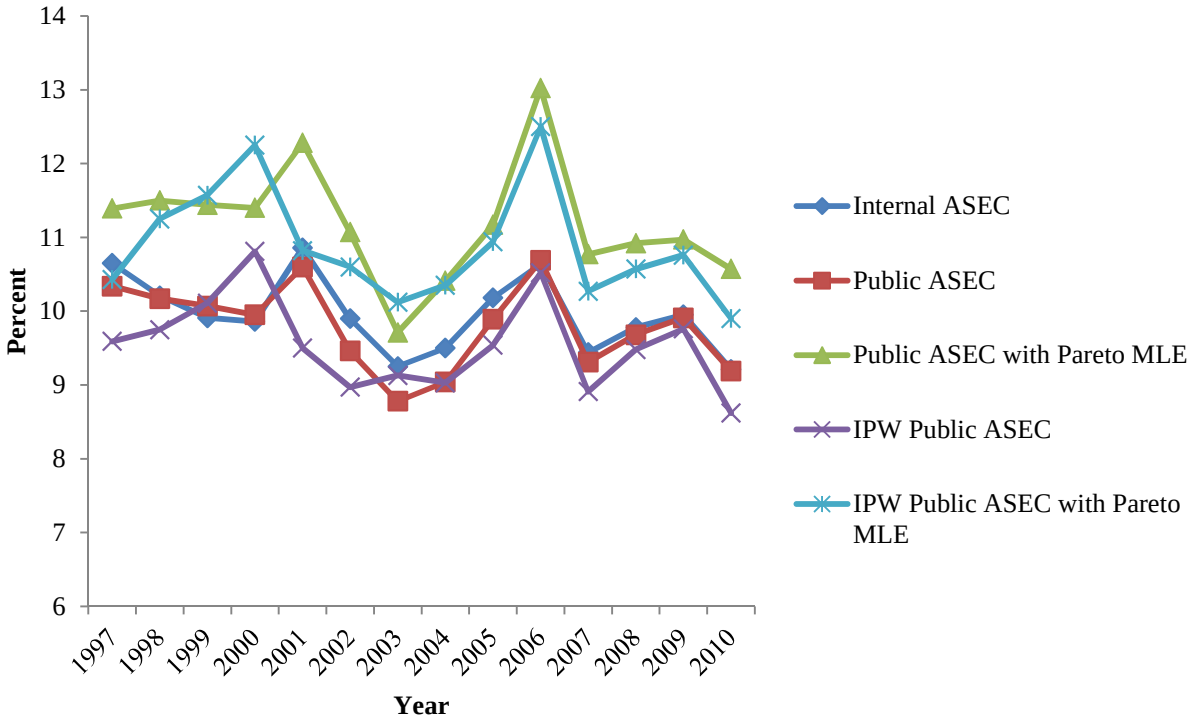

Table 1: Weighted Sample Means of Selected Characteristics of CPS ASEC-DER Matched Sample

Characteristic	Full Sample	Full Time, Full Year	Men	Women
Age	40.71	41.48	40.71	40.71
Gender				
Male (%)	52.20	57.37	100.00	N/A
Female (%)	47.80	42.63	N/A	100.00
Race				
White (%)	83.53	83.56	85.03	81.89
Black (%)	11.11	10.99	9.55	12.81
Other race (%)	5.36	5.45	5.42	5.30
Marital Status				
Married (%)	60.74	63.08	63.19	58.07
Widowed (%)	1.43	1.33	0.59	2.35
Separated or Divorced (%)	13.88	14.19	11.02	17.01
Single, Never-Married (%)	23.94	21.41	25.20	22.57
Educational Attainment				
Less Than High School (%)	8.65	7.44	9.92	7.26
High School Completed (%)	30.72	30.07	31.93	29.41
More than high school (%)	60.63	62.49	58.15	63.33
Hours worked per week	40.58	43.71	43.08	37.59
ASEC Earnings ($2012)				
Respondent	46,311	55,693	56,766	35,180
Nonrespondent	46,206	54,761	56,134	34,492
DER Earnings ($2012)				
Respondent	45,444	55,047	56,358	33,824
Nonrespondent	46,957	55,251	58,334	33,534
Earnings Nonresponse Rate (Item + Whole) (%)	24.56	25.17	25.47	23.57
Top Coded Earnings (%)	6.28	7.84	9.94	2.29
Observations (Unweighted)	897,908	644,463	461,078	436,830

Table 2. Estimated Pareto Shape Parameters from Internal ASEC Earnings (ERN-VAL)

	Full Sample		Full-Time/Full-Year		Men		Women	
Year	Baseline	MLE	Baseline	MLE	Baseline	MLE	Baseline	MLE
1997	1.83	1.89	1.83	1.89	1.82	1.74	1.80	2.18
1998	1.67	1.94	1.69	1.83	1.62	2.01	2.00	2.33
1999	1.81	2.09	1.89	2.37	1.88	2.17	1.71	2.78
2000	1.63	1.84	1.64	1.99	1.66	1.81	1.64	2.51
2001	1.78	2.03	1.68	1.76	1.81	1.99	1.64	2.10
2002	1.81	2.16	1.72	1.87	1.91	2.20	1.59	2.30
2003	2.00	2.17	1.88	1.75	2.05	2.18	1.87	2.08
2004	1.84	2.02	1.85	1.68	1.87	1.86	1.77	2.62
2005	1.73	1.98	1.70	1.72	1.67	1.82	1.97	2.46
2006	1.52	1.91	1.43	1.74	1.36	1.45	1.66	2.42
2007	1.82	2.10	1.76	1.82	1.84	1.98	1.74	2.29
2008	1.74	1.88	1.76	2.00	1.74	1.81	1.83	2.62
2009	1.72	1.75	1.72	1.82	1.75	1.78	1.81	2.29
2010	1.81	2.11	1.95	2.50	1.83	2.37	2.05	2.98
1997-2010	1.76	1.99	1.75	1.91	1.77	1.94	1.79	2.43

Note: The numbers in the table are estimated topcode inflation factors for earnings from longest job in the CPS ASEC (ERN-VAL). The formulas for the baseline and MLE estimates are found in equations (1) and (2) of the paper and Armour et al. (2014)