UKCPR
University of Kentucky
Center for
Poverty Research

Discussion Paper Series
DP 2015-02

ISSN: 1936-9379

Trouble in the Tails? What We Know about Earnings Nonresponse Thirty Years after Lillard, Smith, and Welch

Christopher R. Bollinger
University of Kentucky

Barry T. Hirsch
Georgia State University

Charles M. Hokayem
Centre College

James P. Ziliak
University of Kentucky

September 2015

Preferred citation
Bollinger, C., Hirsch, B., Hokayem, C., Ziliak, J., Trouble in the Tails? What We Know about Earnings Nonresponse Thirty Years After Lillard, Smith, and Welch. *University of Kentucky Center for Poverty Research Discussion Paper Series, DP2015-02.*

Author correspondence
Christopher Bollinger, University of Kentucky, Department of Economics, Gatton College of Business and Economics, Suite 244, Lexington, KY 40506-0034; Email: chris.bollinger@uky.edu Phone: (859) 257-9524

University of Kentucky Center for Poverty Research, Gatton College of Business and Economics, Suite 234, Lexington, KY, 40506-0034
Phone: 859-257-7641; E-mail: ukcpr@uky.edu

EO/AA

Trouble in the Tails? What We Know about Earnings Nonresponse Thirty Years after Lillard, Smith, and Welch

Christopher R. Bollinger, University of Kentucky

Barry T. Hirsch, Georgia State University and IZA, Bonn

Charles M. Hokayem, Centre College

James P. Ziliak, University of Kentucky

September 2015

Abstract: Earnings nonresponse in household surveys is widespread, yet there is limited evidence on whether and how nonresponse bias affects measured earnings. This paper examines the patterns and consequences of nonresponse using internal Current Population Survey individual records linked to administrative Social Security Administrative data on earnings for calendar years 2005-2010. Our findings confirm the conjecture by Lillard, Smith, and Welch (1986) that nonresponse across the earnings distribution is U-shaped. Left-tail "strugglers" and right-tail "stars" are least likely to report earnings. Household surveys understate earnings dispersion, reporting too few low and too few extremely high earners. Throughout much of the earnings distribution nonresponse is ignorable, but there exists trouble in the tails.

Key words: CPS ASEC, nonresponse bias, earnings, measurement error, hot deck imputation

JEL Codes: J31 (Wage Level and Structure)

We thank Adam Bee, Dan Black, Charlie Brown, Bruce Meyer, Chuck Nelson, Trudi Renwick, James Spletzer, and Ed Welniak for helpful comments, plus participants at presentations at the U.S. Census Bureau, Society of Labor Economists, Joint Statistical Meetings, American Economic Association Meetings, Federal Reserve Bank of Cleveland, and university presentations.

1. Introduction

Thirty years ago, Lillard, Smith and Welch (LSW, 1986) brought to the forefront the issue of earnings nonresponse in the Current Population Survey (CPS), providing a sharp critique of Census imputation procedures. Since that time much has changed, some for the better and some not. The Census Bureau responded to the LSW critique and substantially improved the quality of their imputation procedures.[1] Offsetting that progress, however, were sharply rising rates of earnings and income nonresponse in the 1990s and early 2000s. The CPS Annual Social and Economic Supplement (ASEC) and the American Community Survey (ACS) have item nonresponse rates on annual earnings of about 20%. An additional 10% of households that participate in the monthly CPS refuse participation in the ASEC supplement and are assigned the responses of a substitute household (so-called whole imputations).[2] The CPS monthly outgoing rotation group (ORG) files have earnings item nonresponse rates of about 30%.

Individuals for whom earnings are not reported have their earnings "allocated" using hot deck imputation procedures that assign to them the earnings of a "similar" donor who has reported earnings. Although Census imputation procedures improved following LSW, subsequent literature showed that even if nonresponse is missing at random, there exists first-order "match bias" in earnings equation coefficients on variables that are either not included in the Census hot deck such as union status or foreign-born, and on variables that are imperfectly matched such as education.[3] Simple remedies for the bias exist (e.g., omitting imputed earners), but the validity of such remedies, as well as the validity of nearly all imputation methods, relies on the assumption that nonresponse is missing at random (MAR).[4]

Unfortunately, we know surprisingly little about patterns of response bias. LSW (1986, p. 492) speculated that ASEC nonresponse is likely to be highest in the tails of the distribution (U-shaped), but provided no direct evidence since they could not observe earnings for nonrespondents. Given the high earnings nonresponse rates in household surveys, coupled with a paucity of evidence on nonresponse patterns, we address three important and closely-related questions. First, is nonresponse bias ignorable; that is, do respondents and nonrespondents have equivalent earnings, conditional on covariates? This is difficult to know absent external information on nonrespondents' earnings. Second, how do nonresponse

[1] Welniak (1990) documents changes over time in Census hot deck methods for the March CPS.

[2] Unit nonresponse occurs when there is a noninterview or refusal to even the monthly survey. These rates for the basic CPS were between 8 and 9 percent during our sample period (Dixon 2012).

[3] Imputation match bias is discussed subsequently. See Hirsch and Schumacher 2004; Bollinger and Hirsch 2006, and Heckman and LaFontaine 2006.

[4] Following Rubin (1976) and Little and Rubin (2002), we use the term "missing at random" (MAR) to mean earnings data missing at random *conditional* on measured covariates. "Missing completely at random" (CMAR) refers to missingness (nonresponse) not dependent on earnings values, observable or not. Data are "not missing at random" (NMAR) if nonresponse depends on the value of missing earnings, conditional on covariates. We use the term "response bias" (or "non-ignorable response bias") to mean that the earnings data are NMAR.

and patterns of response bias vary across the earnings distribution and are these patterns similar for women and men (or other groups)? And third, can the earnings of survey respondents accurately describe the unobservable distribution of a combined respondent and nonrespondent sample? Answers to these questions have taken on increasing importance in recent years with the expansion of distributional research, whether standard summary measures of unconditional or conditional inequality (e.g. Lemieux 2006; Autor et al. 2008; Burkhauser et al. 2012) or fully specified quantile regression models of earnings (Buchinsky 2012; Kline and Santos 2013).

We address each of the questions above using restricted-access ASEC household files linked to Social Security Administration's Detailed Earnings Records (DER) for March 2006-2011 (corresponding to calendar years 2005-2010). Access to the DER is advantageous as it affords the opportunity to fill in missing earnings for nonrespondents, and to compare survey responses to administrative tax records for respondents. In addition, earnings from the DER are not topcoded, unlike both the public-release and internal versions of ASEC, which improves our estimates of the importance of nonresponse in the right tail of the earnings distribution. We examine whether earnings and response to the earnings question are independent variables. Models of nonresponse as a function of earnings, as measured in the DER, are estimated. We also estimate earnings regressions using the DER, and then examine differences in the residuals from those regressions based on response status. This provides estimates of summary statistics for the conditional distribution of earnings for both respondents and nonrespondents.

In general we find that nonresponse is not ignorable; earnings are not missing at random (NMAR). Respondents and nonrespondents have different earnings distributions. Formally, nonresponse and earnings are not independent, even conditional on a rich set of covariates known to be associated with both variables. We find that – as we allude to in the title – the highest rates of nonresponse are in the tails of the earnings distribution. While on average, male (female) nonrespondents have slightly higher (lower) earnings than respondents, nonresponse is not simply an up or down shift in the distribution. Individuals with earnings that differ substantially from the average (either the gross or conditional mean) are the most likely not to report earnings.

Although it is now understood that use of imputed earnings can introduce serious bias even given the MAR assumption, our finding of NMAR suggests that reliance on respondent samples (even if reweighted) also may provide biased estimates of population earnings. Fortunately, the impact of nonresponse bias on *averages* is small. Using administrative data in which earnings are observed for ASEC nonrespondents as well as respondents, differences in regression coefficient estimates using respondent-only and full samples are typically small. The exception is for coefficients on variables associated with very high or low earnings. In short, the importance of nonresponse is highly dependent on the research question. Given the U-shaped pattern of nonresponse, the nonresponse is particularly important for studies on poverty, earnings and income inequality, and, to a lesser extent, male-female wage differences.

2. Background: Earnings Nonresponse, Imputation Match Bias, and Response Bias

Official government statistics, as well as most research analyzing earnings (and income) differences, include both respondents and imputed earners in their analyses. In the ASEC, earnings nonresponse and imputation rates (we use these terms interchangeably) have increased over time, as stated in the introduction. Figure 1 shows the weighted nonresponse/imputation rates for both earnings and the whole ASEC supplement for March 1988 through 2012 (CY 1987-2011). Item nonresponse for the earnings questions is currently about 20%. There also exists supplement nonresponse and whole imputations. When households participating in the monthly CPS refuse participation in the ASEC supplement, the non-participating households have their entire supplement records replaced by the records from a participating donor household. These whole imputation rates are about 10%. In what follows, we examine separately item nonresponse of earnings and supplement nonresponse.

Researchers typically assume (usually implicitly) that nonresponse does not produce systematic biases in the measurement of earnings. Such an assumption is often unwarranted. For analyses of earnings or wage differentials common in the social sciences, inclusion of workers with imputed earnings frequently causes a large systematic, first-order bias in estimates of wage gaps with respect to wage determinants that are not imputation match criteria or are matched imperfectly in the hot deck procedure. In the appendix, we provide a detailed description of the Census procedure for imputing earnings.

This so-called match bias (see Hirsch and Schumacher 2004; Bollinger and Hirsch 2006; Heckman and LaFontaine 2006) occurs even when nonresponse is missing completely at random. Wage differentials with respect to such attributes as union status, industry, location of residence, foreign-born, etc. are severely attenuated in typical analyses. Estimates using full samples roughly equal the weighted average of largely unbiased estimates from the respondent sample and of severely biased estimates close to zero among the nonrespondent (imputed) sample. For example, the full sample union-nonunion log wage gap estimate for men of 0.142 shown by Bollinger and Hirsch is roughly the weighted average of the 0.191 estimate among earnings respondents and the 0.024 estimate among those with imputed earnings (Bollinger and Hirsch 2006, Table 2). The intuition is simple. Among those for whom earnings are imputed, most union workers are assigned the earnings of a nonunion worker; among nonunion workers, some are assigned the earnings of union workers. Absent a strong correlation between union status and attributes included in the hot deck match, the union-nonunion wage differential in the imputed sample will be close to zero. A more complex bias pattern occurs with respect to the earnings determinants that are included in the hot deck match but grouped into broad categories (e.g., schooling, age, occupation, etc., with gender being the only exact match), leading to imperfect matches between earnings donors and nonrespondents.[5]

Although match bias can be substantial and of first order importance, it is easy to (largely) eliminate. Among the remedies are: exclude imputed earners from the analysis; exclude the imputations and reweight the sample by the inverse probability of response; retain the full sample but adjust estimates using a complex correction formula; or retain the full sample but conduct one's own earnings imputation procedure using all earnings covariates in one's model. In practice, each of these approaches eliminates first-order match bias and produces highly comparable results (Bollinger and Hirsch 2006). Each of these methods, however, assumes earnings are missing at random (MAR); that is, conditional on measured covariates, those who do and do not respond to the earnings questions would exhibit no systematic difference in earnings.[6]

Unfortunately, the validity of the MAR assumption is difficult to test. One approach is estimation of a selection model that directly addresses selection into response rather than assuming MAR (Bollinger and Hirsch 2013). Such an approach generally relies on existence of an exclusion variable(s) that predicts nonresponse but is not correlated with earnings (conditional on controls), as well as reliance on distributional assumptions that cannot be directly verified (though the latter is more readily relaxed with semi-nonparametric selection models). Using CPS survey methods or time period as exclusion variables (these measures affected response rates but not earnings), Bollinger and Hirsch concluded that there exists response bias, with weak negative selection into response (i.e., lower response for those with higher earnings, conditional on covariates). The bias appeared to be larger for men than for women. They found that nonrandom selection bias was largely a fixed effect that showed up in wage equation intercepts, but had little discernible effect on estimated slope coefficients.[7] More problematic, their study (and previous ones) estimates the central tendency for nonresponse bias. As shown in this paper, selection into nonresponse differs across the distribution.

A more direct approach for determining whether or not nonresponse is ignorable, the approach taken in this study, is to conduct a validation study in which one compares CPS household earnings data with administrative data on earnings provided for both CPS earnings respondents and nonrespondents. There are several well-known validation studies comparing earnings information reported in household

[5] Heckman and LaFontaine (2006) provide a correction for standard errors when there are imputations. Stephens and Unayama (2015) extend the match bias concept to IV estimation. If one uses a non-match criterion as an instrument in the first stage (say, state of residence to predict public program receipt), the state coefficient in the first stage is attenuated, leading to upward bias on the "causal" public program coefficient in the second stage.
[6] Note that inclusion of nonrespondents (imputed earners) in the estimation sample, while potentially introducing match bias, does *not* correct for response bias since the donor earnings assigned to nonrespondents are drawn from the sample of respondents. Earnings of nonrespondents are not observed.
[7] This latter conclusion was based on a comparison of wage equation coefficients from their full-sample selection models and those from OLS models in which imputed earners were excluded.

surveys with earnings recorded in administrative data. But typically these studies include only workers reporting earnings in the household survey and do not examine the issue of response bias (e.g., Mellow and Sider 1983; Bound and Krueger 1991; for a survey see Bound, Brown, and Mathiowetz 2001).

We are not the first study to examine response bias using a validation study, but prior studies examining CPS nonresponse are old, use small samples, and examine restricted populations (e.g., married white males). Most similar to our initial analysis is a paper by Greenlees, Reece, and Zieschang (1982), who examine the March 1973 CPS and compare wage and salary earnings the previous year with 1972 linked income tax records. They restrict their analysis to full-time, full-year male heads of households in the private nonagricultural sector whose spouse did not work. Their sample included 5,515 workers, among whom 561 were nonrespondents. Earnings were censored at $50,000. They conclude that nonresponse is not ignorable, with response negatively related to earnings (negative selection into response). Their conclusion is based on a regression of response on administrative earnings, which yields a negative sign, conditioning on a selected number of wage determinants. The authors estimate a wage equation using administrative earnings as the dependent variable for the sample of CPS respondents. Based on these estimates they impute earnings for the CPS nonrespondents. Their imputations understate administrative wage and salary earnings of the nonrespondents by 0.08 log points.[8]

David et al. (1986) conduct a related validation study using the March 1981 CPS linked to 1980 IRS reports. They conclude that the Census hot deck does a reasonably good job predicting earnings as compared to alternative imputation methods. Their results are based on a broader sample and use of a more detailed Census imputation method than was present in Greenlees et al. (1982). David et al. note bias, possibly reflecting negative selection into response.

Although informative and suggestive, it is not known whether results from these early studies examining response bias can be generalized outside their time period and narrow demographic samples. In short, there exists little validation evidence regarding CPS response bias with recent data. With the exception of Kline and Santos (2013), prior studies have not examined differences in response bias across the distribution; the nature of such bias could well differ between the tails and middle of the earnings distribution, as well as between the upper and lower tails. Given the increasing rates of nonresponse over time, it is important to know whether nonresponse is ignorable and, if not, the size and patterns of bias.[9]

Formally the MAR assumption is a statement about the joint distribution, f(Y,R|X), of earnings (Y) and response status (R) conditional on covariates (X). In this case, the covariates we consider are

[8] Herriot and Spiers (1975) earlier reported similar results using these data, the ratio of CPS respondent to IRS earnings being 0.98 and of CPS imputed to IRS earnings being 0.91.

[9] There is a separate literature that considers various methods to deal with missing data. These (very useful) methods, which often require strong distributional assumptions, shed little light on whether CPS earnings nonresponse is ignorable and, if so, how it varies over the distribution.

those used by the Census Bureau in the hot deck imputation procedure and by researchers in estimating models involving earnings. The MAR assumption holds when $f(Y,R|X) = f(Y|X)*f(R|X)$, so earnings and response are independent conditional on covariates, X. It is difficult to summarize this question in the joint distribution, so focus in much of the previous literature (and here) is on the conditional distributions. Formally, the MAR assumption can be tested two different ways using conditional distributions. Does $f(Y|R,X) = f(Y|X)$ and does $f(R|Y,X) = f(R|X)$? Simply put, does the distribution of earnings depend upon the response status or does the response status depend upon the earnings? We examine both of these below. It is important to understand that practical applications may be differently impacted depending upon the particular way in which response and earnings are related. For example, mean earnings regressions (typically estimated with OLS), are only impacted if $E[Y|R,X]$ differs from $E[Y|X]$. The weaker, mean independence condition may hold even when the stronger MAR condition fails (as we will see below).

Greenlees et al. (1982), David et al. (1986) and Little and Rubin (2002) have focused upon $f(R|Y,X)$, the probability of response conditional upon earnings. This provides the simplest and most straightforward test to answer the question of independence. Since R is a binary variable, its entire distribution is summarized by $Pr[R=1|Y,X]$. If earnings (Y) has any predictive power, then earnings and response are not independent, and the MAR assumption fails. Further, the information is useful in a number of contexts. If levels of earnings or income impact response, the relationship is informative for survey design, construction of imputations, and construction of weights to address nonresponse.

It is also informative to examine summary measures of $f(Y|R,X)$, as this is the key distribution for understanding sample selection when Y is the dependent variable in a regression. Unlike $f(R|Y,X)$, the conditional on response distribution of earnings may have multiple parameters (mean, median, quantiles, variance, and skewness for example), which makes it more complex to consider. The classic paper by Heckman (1974) and later papers (for a survey, see Vella, 1998) suggest that a key parameter is either $E[Y|R=1,X]$ or as is often represented, $E[\varepsilon|R=1,X]$, where ε is the error term in a well specified mean regression equation. When the regression of interest is a quantile regression such as the median or other percentiles, it is less clear what the most important parameters will be. We focus upon summary measures of ε from a standard linear regression specification. We also consider differences in the slope coefficient estimates from these regressions.

In the introduction we noted the LSW conjecture that nonresponse is U-shaped with respect to income. In research that was an offshoot of analysis in our paper, Hokayem et al. (forthcoming) use the linked ASEC-DER data to examine how treatment of nonrespondents affects poverty rate estimates. They show a figure with U-shaped nonresponse. Kline and Santos (2013) examine whether returns to schooling and other earnings equation parameters are sensitive to departures from the MAR assumption, using a

1973 March CPS extract linked to IRS earnings data, previously analyzed by Bound and Krueger (1991). They provide evidence that missing data probabilities among men are U-shaped, with very low and high wage men least likely to report. Analysts working with the Consumer Expenditure Survey conclude unit nonresponse is largely uncorrelated with income over most of the distribution, the exception being high nonresponse among very high income households (Sabelhaus et al. 2013).

In section 4 below, we first follow Greenlees et al. (1982) and David et al. (1986) in considering models of f(R|Y). In particular we estimate the probability of response conditional on earnings represented by a set of dummy variables placing earnings in deciles and percentiles. We then turn in section 5 to estimation of models of earnings and examine the distribution of residuals for respondents and nonrespondents across the distribution.

3. Data Description: The ASEC-DER Earnings Link Files

The data used in our analysis are restricted-access CPS ASEC person records linked to Social Security Administration Detailed Earnings Records (DER) for survey years 2006-2011 (reporting earnings for calendar years 2005-2010).[10] In addition to the data included in ASEC public use files, the internal ASEC file has topcoded values for income sources that are substantially higher than the public use top codes.[11]

The DER file is an extract of the Master Earnings File and includes data on total earnings, including wages and salaries and income from self-employment subject to Federal Insurance Contributions Act and/or Self-Employment Contributions Act taxation. Only positive self-employment earnings are reported in the DER because individuals do not make self-employment tax contributions if they have self-employment losses (Nicholas and Wiseman 2009). The DER file contains all earnings reported on a worker's W-2 forms. These earnings are not capped at the Social Security payroll tax 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 capped. This is important given that there are substantial concerns regarding nonresponse and response bias in the right tail of the distribution, but knowledge on these issues is quite limited. That said, in the analysis that follows, we cap DER annual earnings at $2 million to avoid influence from extreme earnings on estimated wage equation coefficients.[12]

[10] The linked ASEC-DER were obtained as part of an internal-to-Census project initiated when one of the authors was a Census employee and analyzed in a secure facility at the U.S. Census Bureau in Suitland, MD. At this time, these data are not available in the wider network of Census Research Data Centers.
[11] Larrimore et al. (2008) document the differences in top code values between the internal and public use CPS files.
[12] Confidentiality agreements under Title 26 of the Internal Revenue Code preclude us from disclosing individual earnings values such as the maximum earnings values in the DER. The two components of our internal ASEC total earnings variable, earnings on the primary job and all other earnings, are each capped at $1.1 million.

The DER file also contains deferred wage (tax) contributions to 401(k), 403(b), 408(k), 457(b), and 501(c) retirement and trust plans, all of which we include in our earnings measure. The DER file does not provide a fully comprehensive measure of gross compensation. Abowd and Stinson (2013) describe parts of gross compensation that may not appear in the DER file such as pre-tax health insurance premiums and education benefits. More relevant for our analysis, particularly for workers in the left tail of the earnings distribution, is that the DER file cannot measure earnings that are off the books and not reported to tax authorities, or earnings that are not subject to Social Security taxation such as primary and secondary teachers in some states. In our analysis, we can compare how discrepancies between ASEC earnings reports (which are likely to include undocumented earnings) and the administrative data change in samples with and without demographic or industry-occupation groups of workers most likely to have undocumented earnings.

Workers in the DER file are uniquely identified by a Protected Identification Key (PIK) assigned by Census. The PIK is a confidentiality-protected version of the Social Security Number (SSN). The Census Bureau's Center for Administrative Records Research and Applications links the DER file to the ASEC via the PIK. Since the Census does not currently ask respondents for a SSN, Census uses its own record linkage software system, the Person Validation System, to assign a SSN.[13] 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. Our examination of ASEC workers not linked to DER indicated that they were disproportionately low wage workers and in occupations where off-the-books earnings are most common. Bond et al. (2013) provide similar evidence using administrative data linked to the American Community Survey (ACS).

Linkage rates between the ASEC and DER administrative data among earners beginning with the 2006 ASEC are about 85 percent. Figure 2 shows the linkage rates across the ASEC wage distribution for both PIK link and the joint PIK and DER linkage. These are useful to look at separately since linkage failures often occur for different reasons. Both rates are lower for those in the left tail of the ASEC wage distribution, but these rates vary little throughout the rest of the distribution.

Since a worker can appear multiple times per year in the DER file if they have several jobs, we collapse the DER file into one earnings observation per worker per year by aggregating total earnings (Box 1 of W-2, labeled "Wages, tips, other compensation") and total deferred contributions across all

[13] The Census Bureau changed its consent protocol to match respondents to administrative data beginning in with the 2006 ASEC (calendar year 2005, the beginning year of our analysis). 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. Beginning with the 2006 ASEC, respondents not wanting to be matched to administrative data had to notify the Census Bureau through the website or use a mail-in response in order to "opt-out". If the Census Bureau doesn't receive this notification, the respondent is assigned a SSN using the Person Validation System. Under the prior "opt-in" consent option in the 2005 ASEC, the linkage rate among earners was 61 percent.

employers. In this way, DER earnings is most compatible with ASEC earnings from all wage and salary jobs (WSAL-VAL), the measure used widely in the labor literature. Like the link to the DER, imputations of earnings occur at the individual level as well. We classify a worker as having imputed earnings if either wage and salary income from the longest job (I-ERNVAL) or from other jobs (I-WSVAL) is imputed. We construct the ASEC and DER average hourly wages by dividing annual ASEC or DER earnings by annual hours worked. Annual hours worked comes from multiplying weeks worked (WKSWORK) by usual hours worked per week (HRSWK).

A measurement issue for workers in some occupations is that workers may report in the ASEC that they received wage and salary earnings, while the company from which they received pay instead reports it to IRS as self-employment earnings. The employer for tax purposes treats them as non-employees (not paying Social Security payroll taxes) and reports earnings on a 1099-MISC in Box 7 ("Nonemployee compensation") rather than a W-2. For example, clergy, real estate agents, and construction workers often receive nonemployee compensation reported to IRS as self-employment earnings, but sometimes report these earnings to the Census as wage and salary earnings.

In line with the labor literature on earnings determination, our goal is to measure wage and salary earnings in a manner similar to that reported in the ASEC, which is influenced by how workers interpret the survey questions on wage and salary earnings. In order to have DER earnings correspond more closely to the ASEC earnings measure, for some workers we include in our DER earnings measure a portion of reported self-employment earnings, with that portion varying by occupation based on the relative frequency of self-employment reports in the ASEC versus DER. Doing so narrows or eliminates what in a few occupations (e.g., the clergy) would otherwise be much lower earnings recorded in DER than in the ASEC. None of our main results is sensitive to this adjustment.[14]

The principal sample used in our analysis includes full-time, full-year, non-student wage and salary workers ages 18 to 65 with positive ASEC and DER earnings reported for the prior calendar year. We do not have information for whole supplement imputations on weeks and hours worked the previous year. Thus, we cannot calculate hourly earnings and must exclude whole imputes from our principal analyses. A separate analysis of the whole supplement imputes is provided in section 6. Our 2006-2011

[14] Specifically, our adjusted DER wage, W^{DER}, is measured by the sum of wage and salary earnings reported in DER, plus some share *r* of self-employment earnings in DER. The occupation-specific adjustment factor *r* is:

r = 1 – (%ASEC-SE / %DER-SE),

where %ASEC-SE and %DER-SE are the respective occupation-specific percentages of ASEC and DER earnings reports that include self-employment earnings. The adjustment factor *r* is zero if ASEC and DER reporting rates of SE are equal, but increases toward 1 as the gap between ASEC and DER SE reporting grows (*r* is set to zero if negative, which occurs if %DER-SE is less than %ASEC-SE). This imperfect adjustment procedure narrows earnings differences between ASEC wage and salary earnings and our DER earnings measure in those occupations where workers often regard themselves as wage and salary workers, but their employers report earnings to IRS as nonemployee compensation.

ASEC-DER regression sample includes 287,704 earners, 157,041 men and 130,663 women. Earnings nonresponse rates (weighted) for this sample is 19.5% among men and 19.3% among women (Table 1).

Table 1 provides summary statistics for our sample by gender. We focus on measures of hourly earnings and earnings response. For women, weighted mean hourly earnings (in 2010 dollars) are equivalent in the ASEC and DER (both $20.80). Among men, ASEC hourly earnings are about a dollar lower than DER earnings ($27.05 versus $28.24), reflecting some very high earnings in DER not fully reflected in the ASEC. Mean log wages are higher in the ASEC than DER, by 0.022 for men and by 0.031 for women.

For men responding in the ASEC, mean DER wages ($28.03) are higher than ASEC wages for these same men ($27.11), but for responding women mean DER wages ($20.91) are equivalent to their ASEC wages ($20.94). For non-responding men, their imputed ASEC wages ($26.81) are substantially lower than their DER wages ($29.14). For non-responding women, the imputed ASEC hourly earnings is an average $20.22, similar to their $20.31 DER wage. Focusing just on DER wages, ASEC male nonrespondents exhibit higher mean DER wages than do respondents ($29.14 versus $28.03), whereas among women nonrespondents exhibit lower DER wages than do respondents ($20.31 versus $20.91). The use of proxy respondents is more prevalent for men than women (53.2% vs. 41.2%).

4. Is Response a Function of Earnings? Nonresponse across the Distribution

Although evidence is highly limited, several previous studies concluded that nonresponse increases with earnings, implying negative selection into response (i.e., as earnings rises, nonresponse increases). Testing this is difficult with public use data since we do not observe earnings for those who fail to respond. We initially follow the approach by Greenlees et al. (1982), who measured the likelihood of ASEC response as a function of linked 1973 administrative (i.e., DER) earnings, conditional on a rich set of covariates, f(R|Y,X). The Greenlees et al. analysis was conducted for white males working full-time/full-year married to non-working spouses.

To examine the conditional distribution of response given earnings, we estimate the following model of nonresponse using our linked ASEC-DER sample:

$$NR_i = \theta lnW_i^{DER} + X_i\beta + u_i. \qquad (1)$$

where the variable NR_i represents an individual i's earnings nonresponse status (0 or 1), lnW_i^{DER} is the natural log of i's DER wage , and X_i includes a rich set of covariates measuring potential experience, race, marital status, citizenship, education, metropolitan area size, occupation, industry, and year. We recognize that this is a relatively simple model of the joint distribution and so, subsequent analysis moves from use of a single linear log wage term to categorical measures for wage percentiles that allow for different responses throughout the earnings distribution. This allows for a less parametric relationship between nonresponse and earnings. Our preferred specification estimates nonresponse rates at each percentile of the earnings distribution, separately for men and women:

$$NR_i = \theta_k Wage\ Percentile_{ik} + X_i\beta + u_i. \qquad (2)$$

Table 2 provides estimates of the nonresponse to earnings relationship using linear probability models, with and without a detailed set of controls, along with the corresponding marginal effects estimates using probit estimation. Because OLS results are highly similar to those from probit, in subsequent tables we show only OLS results. The top panel of Table 2 provides results for men and the middle panel for women. Shown are results with and without controls. Full estimation results on the control variables are available from the authors.

In contrast to Greenlees et al., our results in Table 2 suggests a central tendency of positive rather than negative selection into response.[15] That said, the OLS coefficient for men (with controls) is very close to zero (-0.0225 with s.e. 0.0023), although highly significant given our sample size. Among women, we obtain a larger negative coefficient (-0.0418 with s.e. 0.0027), again indicating that on average nonresponse declines with earnings, conditional on covariates. Absent controls, the R^2 for each regression is effectively zero for men and women, the wage alone accounting for a small fraction of 1 percent of the total individual variation in nonresponse (column 1). Regressions with detailed controls plus the wage account for only 2 percent of the variation (column 3).

Although these results provide what we believe are accurate measures of central tendency for these broad samples of men and women, our results for men appear to be just the opposite of that found by Greenlees et al., who found negative selection into response. Their small sample of married white men with non-working spouses in 1972, however, is not representative of today's workforce. In order to compare our results with those of Greenlees et al., we create a roughly similar sample restricted to married white male citizens with spouse present. Unlike Greenlees et al., we include those with working spouses since married women's labor force participation is now closer to the norm rather than the exception. We refer to this as our "Mad Men" sample, shown in the bottom panel of Table 2. This sample is likely to have a small proportion of workers in the far left tail of the DER earnings distribution. In contrast to the negative coefficients on log earnings of -0.0268 and -0.0225 for all full-time/full-year men (columns 1 and 3), using the Mad Men sample flips the signs and produces coefficients of 0.0193 and 0.0122 (with s.e. of 0.0028 and 0.0035). The latter results are qualitatively consistent with Greenlees et al., as well as previous studies finding negative selection into response, though again we emphasize that the "Mad Men" sample is restrictive and not representative of the modern labor force.

[15] Greenless et al estimate the probability of response, while we estimate the probability of nonresponse. Greenlees et al. find a negative coefficient, indicating less likely to respond, while our results indicate more likely to respond.

Rather than focusing on central tendency, it is more informative to examine how nonresponse varies across the distribution. To examine whether nonresponse changes across the distribution, we initially modify the nonresponse equation specification by grouping the bottom 90% of earners into wage deciles, while breaking up the top decile into finer percentile increments. As seen in Table 3, ASEC nonresponse regressions are estimated for men, women, and the 'Mad Men' sample, with DER wage decile and percentile dummies included, with and without controls (the intercept is suppressed). Each decile/percentile coefficient represents the nonresponse rate at the given DER wage level. Readily evident from the coefficients is that nonresponse rates are not constant across the distribution. Rather, there exist U-shaped patterns of nonresponse, as hypothesized by Lillard et al. (1986). Focusing first on the male equation with controls (column 2), nonresponse is particularly high in the first decile of the DER wage distribution (0.192), roughly double the level seen throughout most of the distribution. Nonresponse then rises sharply at the top percentiles. The central tendency of positive selection into response among men in Table 2 appears driven by the high nonresponse rates in the left tail of the distribution.

Women exhibit a similar but weaker U-shaped pattern of nonresponse than do men. Their unconditioned nonresponse rates across the deciles are similar to those for men (column 3 versus 1). Women differ from men in that their conditional rates of nonresponse are lower (column 4 versus 2) and they do not exhibit as large of increases at the top percentiles. Note that the percentiles for women and men differ, the wage at the higher percentiles for women being substantially lower than for men. Below we examine nonresponse rates across percentiles of a common joint wage distribution for men and women. Also evident from Table 3 is that the 'Mad Men' sample of married white male citizens does not exhibit so strong a U-shaped nonresponse pattern in the left tail as does the larger population of women or men. Rather, we observe relatively flat nonresponse throughout much of the earnings distribution before exhibiting rising nonresponse in the top percentiles.

Patterns of nonresponse across the entire distribution are most easily discerned visually. In Figure 3, we show nonresponse rates for both men (top panel) and women (bottom panel) for each percentile of the DER wage distribution. The top curve in each panel shows the weighted mean rate of nonresponse at each percentile of the DER wage distribution, absent covariates. The lower curve for each is based on equation (2), which includes a large set of covariates and a full set of percentile dummies (with one omitted percentile). We follow Suits (1984) and adjust the values of all the percentile dummy coefficients (along with the "zero" omitted percentile) to provide a measure of the conditional nonresponse rate at each percentile, relative to the mean rate.[16] By construction, the 100 values shown in the lower curve sum

[16] The Suits (1984, p. 178) adjustment factor is the value k that makes the average of the percentile coefficients equal to zero. That is, $k = -(b_2 + b_3 + \ldots b_{100} + 0)/100$, where b represents the 99 included percentile dummies. The value k is added to each b and to "zero" for the omitted percentile. These Suits-adjusted coefficients are shown in the lower curves in Figure 3.

to zero.

The pattern in Figure 3 for men shows a U-shape, with considerably higher nonresponse in the lower and upper tails of the distribution, but with rather constant nonresponse rates from about the 20th through 95th percentiles. There is very little difference between the unadjusted (top) and adjusted (bottom) curves, apart from the downward adjustment of the latter to reflect measurement relative to the conditional mean rate. Whereas we see nonresponse decline in the left tail throughout much of the first quintile, rising nonresponse is restricted to the top ventile. Nonresponse is largely uncorrelated with the wage throughout most of the distribution, the obvious exceptions being in the tails of the distribution. (i.e. "trouble in the tails"). Figure 3 provides strong support for the LSW U-shape conjecture.

The evidence for women (lower half of Figure 3) is qualitatively no different from that seen for men, indicating a U-shaped nonresponse pattern. That said, there are some differences in the magnitudes of the tails. In the lower-end of the wage distribution, women exhibit slightly higher rates of adjusted and unadjusted nonresponse than do men. In the right tail of the distribution, however, women exhibit minimal increases in nonresponse, increases not easily discerned until one moves to the highest percentile. Although referring to the nonresponse pattern as "U-shaped" is convenient shorthand, emphasis should be given to the high rates of female nonresponse in the left tail of the distribution coupled with rather similar rates throughout the rest of the distribution outside of the very top percentile.[17]

The male and female nonresponse curves shown across the wage distribution in Figure 3 are based on the gender-specific wage percentiles. At a given percentile, say the 90th percentile, the wage for men will be considerably higher than that for women. In Figure 4, we form percentiles based on the joint male-female DER wage distribution and then show the unadjusted nonresponse rates for men and women at each percentile of this common distribution. The male and female curves shown in Figure 4 are remarkably similar, indicating that women and men have similar likelihood of nonresponse at similar wage levels. We saw previously that high nonresponse in the left tail is more evident among women and high nonresponse in the right tail is most evident among men. These patterns appear because women are disproportionately concentrated in the left tail and men in the right tail. With a joint earnings distribution, male and female nonresponse behaviors are highly similar when compared at the same wage levels.

Our final evidence in this section is to show nonresponse rates for men and women with respect to percentiles across the *predicted* wage distribution, seen in Figure 5. Although this does not test for

[17] Coefficients on control variables in the nonresponse equations (available on request) provide information on which types of workers are least and most likely to respond to the ASEC earnings questions, conditional on the wage (using the full set of percentile dummies). For the most part, demographic, location, and job-related measures account for little of the variation in response. Coefficients are generally similar for men and women. Most notable are high nonresponse probabilities found among workers who are black, Asian, never married, and residents in large (5 million plus) metro areas. Public sector workers are more likely to report earnings.

response bias, the results are informative, showing how nonresponse is related to an index of earnings attributes (education, demographics, location, and job type). The predicted DER wage for each worker is calculated based on coefficient estimates and worker attributes from the earnings equation

$$lnW_i^{DER} = X_i\beta + \epsilon_i \quad (3)$$

We use the same samples of linked ASEC-DER respondents and nonrespondents and same set of covariates used in the previous nonresponse equations, but with the log wage shifted to the left-side of the regressions. In addition to showing how nonresponse varies with each percentile of the predicted wage, we also show an OLS line fitted to the nonresponse points.

For women, Figure 5 provides little evidence of high nonresponse in either tail of the attribute distribution, let alone a U-shape. Men exhibit somewhat higher nonresponse in the left tail of the index and a slight rise in the right tail. For the most part, nonresponse for men and women is fairly constant throughout the attribute distribution, with a gradual decline in nonresponse as earnings attributes increase. What accounts for the U-shaped nonresponse (i.e., trouble in the tails) is not earnings attributes per se; rather, it is the *realization* of either very low or very high earnings.

Our interpretation of the nonresponse evidence up to this point is straightforward. The good news is that earnings nonresponse in the ASEC appears to be largely ignorable throughout much of the earnings distribution, varying little with the realized level of earnings, conditional on covariates. To the extent that there is a pattern over the 10th to 95th percentiles, it is one of nonresponse declining ever so slightly with respect to earnings over the distribution before turning up at the very top percentiles. We regard any such pattern between the 10th and 95th percentiles as inconsequential. Where there most clearly exist problems is in the tails. Stated simply, nonresponse is highest among “strugglers” and “stars”. Characterizing selection into response based solely on estimates of central tendency over entire distributions, as seen in Table 2 and in prior literature, is largely uninformative and potentially misleading.

The analysis in this section has identified the pattern of response bias. It is difficult to provide direct evidence on the causes of U-shaped nonresponse, given that we have already conditioned on a rich set of covariates. Plausible explanations, however, can be offered. High rates of nonresponse in the very top percentiles of the distribution are likely to stem from concerns about confidentiality or the belief that there is no compelling duty to report one’s earnings to a government survey agency. These percentiles roughly correspond to where individual earnings are top coded in public use ASEC files. Analysis of workers with topcoded earnings is already difficult for researchers using public files; high nonresponse among such earners makes such analysis all the more difficult.[18]

[18] Researchers using the CPS often assign mean earnings above the topcode based on information provided by Census or by researchers using protected internal ASEC files (Larrimore et al. 2008). Because very high earners are

High rates of nonresponse among those with low earnings (conditional on covariates) may stem from several reasons. Discussions with Census field representatives suggest that some ASEC participants find it difficult to report annual earnings and income measures, despite attempts to help them produce such information (e.g., prompts regarding the amount and frequency of typical paychecks). Substantial effort may be required for many low-income household members to report earnings; these high effort costs decrease response. Consistent with this explanation, Kassenboehmer et al. (2015) examine paradata measuring the fraction of survey questions answered in an Australian household survey. The authors conclude that nonresponse for income and other "difficult" questions results in part from cognitive difficulties in answering such questions, based on evidence that a "fraction answered" variable behaves statistically much like a cognitive ability measure in the relationship between education and earnings.

An additional explanation offered by persons knowledgeable about the ASEC is that high rates of nonresponse for earnings and other income sources, particularly among low-wage women, may result in part from the (invalid) concern that reporting such information to Census might place income support program eligibility at risk. Finally, it is worth noting that some of the nonresponse in the left tail of the earning distribution might be associated with off-the-books earnings. Workers likely to have off-the-book earnings, which leads to lower DER earnings, may also be less likely to answer ASEC earnings questions. This is an issue we examine subsequently, finding that the omission of workers in occupations where off-the-books earnings are most common has little discernable effect on the pattern of nonresponse.

5. Is Earnings a Function of Response? DER Wage Residuals across the Distribution

The prior section provided evidence on the distribution of response conditional on earnings and other covariates. In this section we examine the distribution of earnings conditional on response and earnings covariates, f(Y|R,X). We estimate wage regressions specified in equation (3) using lnW_i^{DER} and provide kernel density estimates of residuals for respondents and non-respondents. We also provide summary statistics of the residuals by response status and test for differences between the two distributions.

The estimated distributions of residuals are presented in Figure 6. The top panel of Figure 6 presents the distributions by response status among men, while the bottom panel does so for women. In both panels, peaks of the respondent distribution are higher than peaks of the nonrespondent distributions. Similarly, the tails of the nonrespondent distribution are generally longer, indicating a higher variance for nonrespondents. Table 4 supports this, demonstrating that the variance for male (female) nonrespondents is 1.63 (1.72) times the variance of male (female) respondents. Testing differences between these variances using either the standard F-test or Levine's test rejects the null hypothesis of equivalence at

less likely to report earnings in the ASEC, there will be some understatement of high-end earnings due to non-ignorable response bias. An implication from our research is that topcode multiples should be somewhat higher than those recommended based on the estimated mean earnings of ASEC respondents above the topcode.

conventional levels. Tests for differences in means reject the null hypothesis as well. A simple test of the difference in the medians fails to reject for men, but does reject for women. Examining the percentiles shows the major differences occur in the tails, as seen in Figure 6. We conclude that there is strong evidence of differences between these distributions, with the most substantive differences in the variances and other higher moments. Furthermore, Kolmogorov-Smirnov tests reject the null that f(Y|Respondents,X) = f(Y|Nonrespondents,X), which is a sufficient condition to reject the hypothesis that f(Y|R,X) = f(Y|X). In Section 7, which provides guidance for users of publicly available CPS files, we consider the practical importance of these differences.

Information in Table 4 provides an alternative way to "view" the pattern of residuals across the distribution. The column NR-R shows differences in DER wage residuals between ASEC nonrespondents and respondents at selected percentiles. For men and women, NR-R differences change from highly negative to highly positive as earnings increase. In the left tail we see positive selection into response, with ASEC nonrespondents having lower DER earnings residuals than respondents. In the middle of the distribution, NR-R differences are close to zero, indicating little response bias. At the top of the distribution, positive NR-R residual differences indicate negative selection into response.

6. Additional Evidence and Robustness Checks

In this section, we provide evidence and robustness checks complementary to our prior analysis. We examine (a) DER earnings among households who did not participate in the ASEC supplement (whole imputations); (b) how the sample exclusion of students and those who do not work full-time/full-year affected results; (c) identification of occupations and worker groups with relatively large shares of earnings off-the books (i.e., not recorded in DER) earnings; and (d) a robustness check in which our estimation sample is rebalanced to reflect underrepresentation of certain types of workers and jobs due to failure to create ASEC-DER linkages, either because an individual PIK is absent or the PIK cannot be linked in DER records.[19]

Whole imputations. As seen previously in Figure 1, roughly 10 percent of households who participate in the basic CPS refuse to participate in the ASEC supplement. A non-participating household is then assigned ASEC values based on a "whole impute" from a participating donor household.

[19] In analysis not reported in this paper, we examine the accuracy of proxy responses, which account for roughly half of CPS earnings reports, as seen in Table 1. Moreover, earnings nonresponse is substantially higher among individuals with a proxy respondent (Bollinger and Hirsch 2013). If one includes proxy dummies in a standard CPS wage equation, one finds substantive negative coefficients associated with the use of non-spouse proxies and coefficients close to zero for spousal proxies. Our unreported analysis with the linked ASEC-DER data indicates that both spouse and non-spouse proxy reports are accurate, the exception being modest underreporting of married men's earnings by wife proxies (for related evidence, see Reynolds and Wenger 2012). The substantive proxy wage effects found in a standard Mincerian wage equation do not reflect misreporting, but instead worker heterogeneity correlated with proxy use not captured by standard covariates.

Households with whole imputes are excluded from our main analysis. We do observe DER earnings for the original nonrespondent household, plus information for these individuals reported in the monthly CPS survey. Absent supplement information on weeks and hours worked the previous year among nonparticipants, however, we are unable to calculate an hourly wage measure. Instead we focus on annual earnings.

In Figure 7, we show how whole supplement nonresponse differs across the joint DER earnings distribution among men and women, similar to the approach seen in Figure 4 for item nonresponse. Note that the sample here differs from that seen elsewhere in the paper. Because we cannot observe full-time and full-year status among supplement nonparticipants, we define our sample as those who were full-time workers during the March survey reference week. Hence, the samples of both supplement participants and nonparticipants will include mostly FT/FY workers coupled with a small share of workers not FT/FY. We also examined supplement nonresponse patterns for men and women conditioned on measured attributes (as seen in the Suits measure in Figure 3). Here we controlled for a rich set of covariates available in the monthly CPS, these measures being available for supplement participants and nonparticipants. These results are available upon request.

Figure 7 shows a clear-cut pattern. Supplement nonresponse (a form of unit nonresponse) is highest among those with low earnings, with a gradual and slight decline as earnings increase. In short, there is positive selection into supplement participation, with a disproportionate share of low earnings workers not participating. We observe little difference in supplement nonresponse between men and women, as evident by comparison of the squares and diamonds in Figure 7. Although supplement non-participation is lower than is item nonresponse for earnings (roughly 10 versus 20 percent), positive selection into supplement participation likely leads to a modest understatement of both earnings inequality and poverty (Hokayem et al. Forthcoming).

Sample exclusions. Excluded from our sample were students and those who did not work full time/full year. The purpose of the exclusion was to help us focus on a population that has a relatively strong attachment to the labor market. School attendance questions are asked only of those below age 25. Although a considerable number of young persons are excluded, the total number is not large. The sample of workers who did not work full year or full time per week is a more substantive share of the sample. As a robustness check, we examine whether the nonresponse pattern for these excluded workers is similar to that seen for our primary sample. Figure 8 shows nonresponse rates for these excluded workers, by gender, at each percentile of their respective DER wage distribution. The pattern of nonresponse is noisy, as expected given their relatively small sample sizes. But both men and women display slightly U-shaped patterns of nonresponse, similar to those of our main samples. In contrast to results from our primary samples, one does not see extremely high rates of nonresponse in the lower tail or at the highest percentiles among students and workers who are not FT/FY. That said, inclusion of these workers in the main sample does not alter our principal results in any substantive way.

Occupations with off-the-books earnings. To gather information on workers who cannot be linked to tax records or have earnings off-the-books, we identify (a) occupations in which many ASEC workers cannot be linked to DER wages (Table 5) and (b) occupations in which there are large gaps between earnings reported in the ASEC and earnings reported in DER administrative records (Table 6).

Recall that overall linkage rates of the ASEC sample to DER are about 85% (see Figure 2). The top half of Table 5 lists occupations with the lowest rates of a link of ASEC earners to a PIK number; the bottom half provides the linkage rate to DER earnings. Note that the DER and PIK linkage rates are based on the same denominator of ASEC earners. For example, among the sample of 2758 construction laborers, 1884 are linked to a PIK (68.3%), as reported in the top half of Table 5. Of those 1884 workers, 1710 (90.8%) are linked to DER earnings, producing an overall DER linkage rate (seen in the bottom half of Table 5) of 62.0% (1710 out of 2758). Among the occupations with low PIK and DER linkages are the construction trades (e.g., painters, drywall installers, roofers, brick masons, laborers, and helpers); dishwashers, cooks, dining attendants and bartender helpers, and food preparation workers; grounds maintenance workers; and agricultural and fishing related workers.

Using our linked ASEC-DER sample, we also examine which occupations show the largest percentage (log) gap between ASEC earnings and reported DER earnings. These occupations are shown in Table 6. Not surprisingly, there is overlap between the occupations listed in Tables 5 and 6. Occupations including jobs with workers and/or earnings off-the-books also have workers for whom some portion of earnings is reported and some is not. In addition to the types of occupations summarized above, we see large ASEC minus DER earnings gaps for occupations such as door-to-door sales workers, real estate brokers and agents, bartenders, and workers in construction trades. A simple way to characterize "high-gap" occupations is that they include jobs or types of work where there is often an opportunity to avoid reporting earnings (Roemer 2002). In addition, many of these occupations are ones in which earnings (or some share of earnings) are reported to IRS and SSA as self-employment earnings, but that household members may report to Census as wage and salary earnings. Recall that our DER earnings measure includes a share of self-employment earnings, with that share varying by occupation.[20]

How serious is off-the-book earnings for our analysis? It appears to be far less of a problem than expected. Our concern was that a substantive portion of the high nonresponse seen in the left tail of the DER wage distributions was the result of workers with off the book earnings being less likely to answer

[20] Absent that adjustment, the clergy was high on the list for gaps between ASEC and DER earnings. With the adjustment, the gap is close to zero. Clergy are typically taxed as self-employed workers, but often report earnings in the ASEC as wage and salary earnings, creating a large gap between ASEC and DER earnings.

ASEC earnings questions. Similarly, the negative values of DER wage residuals for nonrespondents minus respondents (NR-R) seen in the left tail of the distribution (Table 4), which we have interpreted as positive selection into response among low wage earners, might partly reflect underreported earnings. Although we cannot rule out such concerns, our robustness checks suggest this is not a serious problem. In Figure 9, we remove from our male and female samples all workers in the "high gap" occupations included in Table 6, and then additionally remove all foreign-born noncitizens, some of whom may have high rates of earnings off the books. For both men and women, what we find is an almost total overlap in nonresponse rates in the left tail (and elsewhere) for the full and restricted samples.[21]

Rebalancing. As previously documented, our estimation sample does not include ASEC participants who could not be linked to DER earnings records. As a robustness check, we reweighted our sample using inverse probability weighting (IPW), attaching higher weight to individuals with characteristics associated with low probabilities of a link, and lower weight to those with characteristics associated with high linkage probabilities. Probabilities were estimated using probit estimation modeling DER linkages as a function of demographic and location attributes, plus detailed occupation and industry dummies. We then created the weighted nonresponse figures across the distribution, equivalent to those seen previously in Figure 3. The rebalanced IPW figures effectively overlay those shown in Figure 3, with no discernable differences. Absent visual differences, we do not show these rebalanced IPW figures.

7. Dealing with Nonresponse: Guidance for CPS Users

The analysis in this paper has implications for researchers using the CPS, as well as similar household data sets such as the American Community Survey (ACS). As emphasized in previous work, discussed earlier in the paper, even if nonresponse were completely missing at random, severe "match bias" can arise in the estimation of earnings equation coefficients if researchers include nonrespondents with earnings imputed by Census. Attenuation is severe for coefficients on variables not used in the earnings hot deck match. Bias is more complex when earnings have been imputed using an imperfect match of donor attributes (e.g., schooling, age, etc.). Among the several "remedies" for match bias (Bollinger and Hirsch 2006), the simplest and most widely used is to throw out imputed earnings and rely on analysis with only respondents. The respondent sample can be reweighted by the inverse probability of response, but in practice this typically makes little difference.

The linked ASEC-DER data allow us to examine directly whether relying solely on respondents' earnings produces results similar to what would be produced using complete (but unobtainable) data. Because the DER sample includes administrative earnings for nonrespondents as well as respondents, we

[21] Foreign born noncitizens are disproportionately employed in occupations with high levels of off-the-books earnings. As compared to native men and women, however, our nonresponse equations show that the conditional rates of earnings nonresponse are lower among foreign born noncitizens.

can compare earnings function parameter estimates from respondent-only samples with those from complete samples, something not possible with publicly-available data.

Using the DER sample, we estimate log wage equations with a dense set of covariates, separately for the respondent, nonrespondent, and pooled samples. Using estimates from these regressions, in Table 7 we provide the predicted wage for men and women using means from the full ASEC sample multiplied by coefficient estimates from the regressions using the alternative samples. We use as our benchmark the predicted earnings based on coefficients from the full sample, not obtainable using ASEC data because of the absence of nonrespondents' earnings. We compare the full-sample predicted wage to those obtained using the coefficients from the respondent sample, which can be calculated using public ASEC data.

Focusing first on men, use of full sample coefficients with the full sample worker attributes (X's) results in a predicted mean log wage of 3.066. This is close to that obtained using respondent-only betas, which leads to a predicted mean log wage of 3.075, or 0.009 (one percent) higher than obtained with the full sample. The equivalent values for women are 2.809 using full sample betas and 2.825 using respondent betas, a 0.016 difference. An implication of this result is that our inability to observe earnings for nonrespondents causes gender wage gaps to be understated by less than 1 percentage point (the differences between 0.009 for men and 0.016 for women). As seen earlier, these differences reflect the mean tendency toward positive selection into response in the ASEC, more so for women than men. Selection is more readily evident comparing predicted earnings using respondent (R) and nonrespondent (NR) betas. The R–NR predicted earnings difference is 3.075–3.035 = 0.040 for men and 2.825–2.744 = 0.081 for women. These differences are substantive. Because the nonrespondent shares of the total samples are relatively small (roughly 20 percent), the respondent only sample provides coefficient estimates close to what would be produced using the full sample, the latter not being an option with public use data.

Although our assessment regarding the reliability of respondent-only samples is very much a positive one, this assessment is based on the accuracy of mean outcomes. As seen in the paper, the news is less rosy in the tails. Bias from nonresponse prevents researchers from observing many low earners over a fairly wide range and many high earners at the very top of the distribution. The former may be the more serious problem, at least for researchers using public use data. High nonresponse in the lower tail affects our ability to measure and understand low wage labor markets, low income households, and poverty. Problems in the right tail are concentrated among the very top percentiles, where individuals already have their earnings masked (topcoded) in public use files. Research on very high earners is severely constrained, even absent nonresponse. That said, public use files no doubt include too few topcoded earners due to response bias.

As noted above, mean differences between the full and respondent samples are quite small. Substantive differences between the two distributions are in spread and higher moments. Sample selection is predominantly a function of location. To examine this, we assess the reliability of respondent-only coefficients by estimating models using the DER wage for both the respondent-only and full samples. We can then establish whether the respondent sample produces coefficients close to those from the full sample. We argue that a full sample is the "preferred" sample frame from which to construct estimates, comparable in principle to estimates that could be obtained from the ASEC had nonrespondents responded. Table 8 provides selected coefficients for men and women, respectively. Full regression results are available from the authors. We attach the symbol Ψ to each full sample coefficient that is significantly different (at the 0.05 level) from that obtained with the respondent-only sample.

The results in Table 8 are very much in line with prior evidence. For variables most associated with low earnings (e.g., black and foreign born) the unbiased "All" DER coefficients for men are moderately (but significantly) lower than the respondent-only betas, consistent with the respondent sample failing to include a large number of low-wage workers in the far left-tail. For variables associated with very high earnings (MA, professional, and PhD degrees) the unbiased All DER samples for men and women produce coefficients larger than does the respondent-only sample, the latter failing to include workers in the far right tail. For variables associated with middle-range earnings (HS, some college, and associate degrees) the two sets of coefficients are neither statistically nor quantitatively different.

Evidence from Table 8, coupled with prior evidence in the paper, leads to the conclusion that the importance of nonresponse (selection into the CPS and other household surveys) is dependent upon the research question. If one is interested in central tendency or interested in the impact of wage attributes associated with earnings toward the middle of the wage distribution, use of respondent-only samples are likely to provide coefficient estimates with trivial bias. For wage attributes associated with extremely low or high earnings, a respondent-only sample will understate to some modest degree the absolute value of the coefficient (being insufficiently negative in the former and too positive in the latter). Neither MAR nor a weaker definition of MAR, E[Y|X,R]=E[Y|X], strictly hold. It is important to recognize, however, that failure of MAR also implies that the hot deck imputation procedure adopted by Census, as well as other imputation procedures, are invalid as well. Imputations rely heavily upon MAR and are not a remedy for selective nonresponse. Fortunately, respondent-only samples do "reasonably well" for most (but not all) analyses. Inclusion of imputed earners in one's analysis does not solve the selection problem since nonrespondents are assigned earnings of donors who do respond. At the same time, inclusion of imputed earners often introduces severe match bias on coefficients on attributes not included in (or matched imperfectly) in the Census hot deck procedure.

8. Conclusion

This paper set out to address three questions not adequately examined in prior literature using a unique restricted-access dataset that links ASEC household files to administrative earnings records. First, is nonresponse ignorable; that is, do respondents and nonrespondents have equivalent earnings, conditional on covariates? Throughout much of the distribution there is little correlation between response and earnings, suggesting nonresponse is largely ignorable over this range, but with possible trouble in the tails. Second, how do nonresponse and patterns of response bias vary across the earnings distribution and are these patterns similar for women and men (or other groups)? Nonresponse across the earnings distribution, conditional on covariates, is U-shaped, with left-tail “strugglers” and right-tail “stars” being least likely to report earnings. Women have particularly high nonresponse in the left tail; men have high nonresponse in the far right tail. Using a joint distribution of wages, we see little difference between men and women in nonresponse at the same wage levels. Third, can the earnings of survey respondents accurately describe the unobservable distribution of a combined respondent and nonrespondent sample? The reliability of conclusions drawn from a respondent-only sample is generally very good, and often the best one can do. Because those with unusually low and high earnings, conditional on measured attributes, are disproportionately missing from the sample, wage equation coefficient estimates on attributes associated with very low (high) earnings will be understated in absolute value. Gender wage gaps are slightly understated. More broadly, research on low-wage workers and on top earners, as well as on earnings and income inequality, is likely to be affected by nonignorable response bias.

References

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

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

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

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

Bond, Brittany, J. David Brown, Adela Luque, and Amy O'Hara. "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, 2013.

Bound, John, Charles Brown, and Nancy Mathiowetz. "Measurement Error in Survey Data," in *Handbook of Econometrics, Vol. 5*, edited by E. E. Leamer and J. J. Heckman, Amsterdam: Elsevier, 2001, 3705-3843.

Bound, John and Alan B. Krueger. "The Extent of Measurement Error in Longitudinal Earnings Data: Do Two Wrongs Make a Right?" *Journal of Labor Economics* 9 (1991): 1-24.

Buchinsky, Moshe. "Changes in the U.S. Wage Structure 1963-1987: Application of Quantile Regression." *Econometrica* 62(2), 2012: 405-458.

Burkhauser, Richard V., Shuaizhang Feng, Stephen Jenkins and Jeff Larrimore. "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) (May 2012): 371-388

David, Martin, Roderick J. A. Little, Michael E. Samuhel, and Robert K. Triest. "Alternative Methods for CPS Income Imputation," *Journal of the American Statistical Association* 81 (March 1986): 29-41.

Dixon, John. "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, (August 2012): 1977-1982.

Fitzgerald, John, Peter Gottschalk, and Robert Moffitt. "The Michigan Panel Study of Income Dynamics," *Journal of Human Resources* 33 (Spring 1998): 251-299.

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

Heckman, James J. "Shadow Prices, Market Wages, and Labor Supply," *Econometrica* 42 (July 1974): 679-694.

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

Herriot, R. A. and E. F. Spiers. “Measuring the Impact on Income Statistics of Reporting Differences between the Current Population Survey and Administrative Sources,” *Proceedings*, *American Statistical Association Social Statistics Section* (1975): 147-158.

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

Hokayem, Charles, Christopher Bollinger, and James P. Ziliak. “The Role of CPS Nonresponse in the Measurement of Poverty,” *Journal of the American Statistical Association* (forthcoming).

Kassenboehmer, Sonja C., Stefanie Schurer, and Felix Leung “Testing the Validity of Item Nonresponse as a Proxy for Cognitive and Non-Cognitive Skills,” IZA DP No. 8874., February 2015

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

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

Larrimore, Jeff, Richard V. Burkhauser, Shuaizhang Feng, and Laura Zayatz. “Consistent Cell Means for Topcoded Incomes in the Public Use March CPS (1976-2007).” *Journal of Economic and Social Measurement* 33 (2008): 89-128.

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

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

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

MaCurdy, Thomas, Thomas Mroz, and R. Mark Gritz. “An Evaluation of the National Longitudinal Survey of Youth,” *Journal of Human Resources* 33 (Spring 1998): 345-436.

Mellow, Wesley and Hal Sider. “Accuracy of Response in Labor Market Surveys: Evidence and Implications,” *Journal of Labor Economics* 1 (October 1983): 331-344.

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

Reynolds, Jeremy and Jeffrey B. Wenger. "He Said, She Said: The Gender Wage Gap According to Self and Proxy Reports in the Current Population Survey," *Social Science Research* 41 (March 2012): 392-411.

Roemer, Mark. “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, US Census Bureau, 2002.

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

Sabelhaus, John, David Johnson, Stephen Ash, Thesia Garner, John Greenlees, Steve Henderson,and David Swanson. “Is the Consumer Expenditure Survey Representative by Income?” NBER Working Paper No. 19589, October 2013.

Stephens Jr., Melvin and Takashi Unayama. 2015 “Estimating the Impacts of Program Benefits: Using

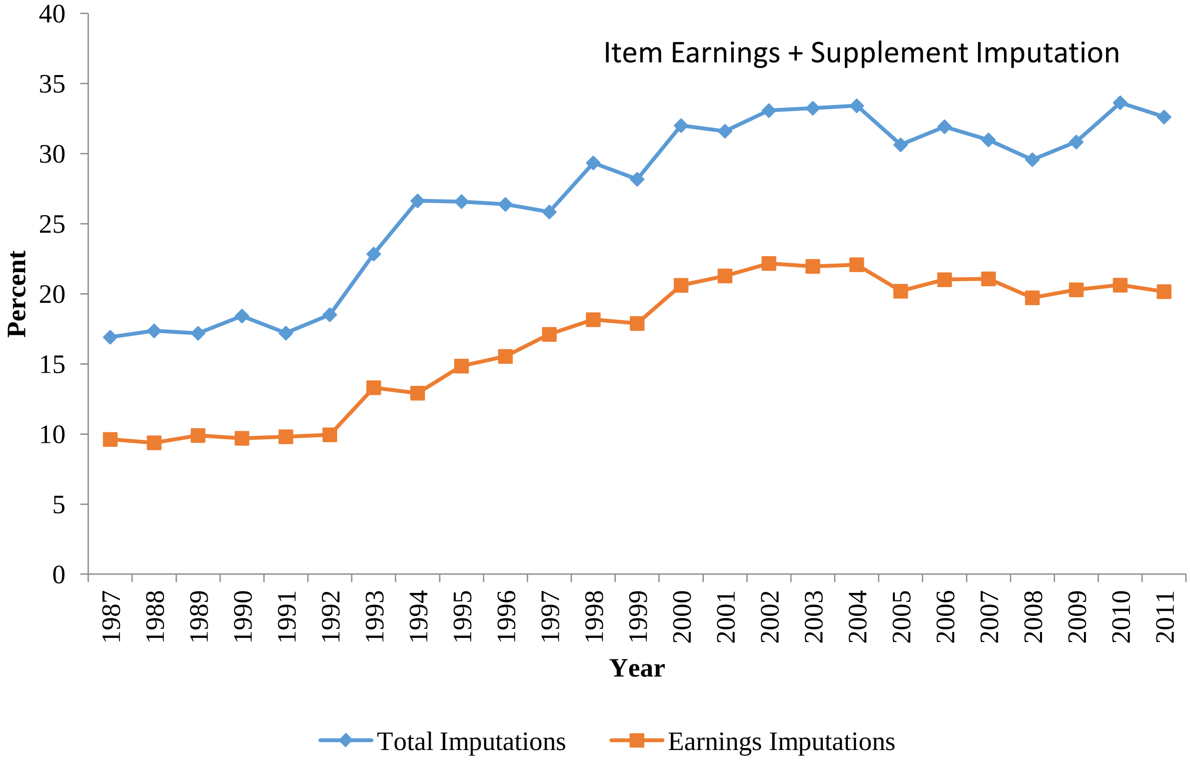
Instrumental Variables with Underreported and Imputed Data," NBER Working Paper 21248, June.

Suits, Daniel B. "Dummy Variables: Mechanics V. Interpretation," *The Review of Economics and Statistics* 66 (February 1984): 177-180.

Vella, Francis. "Estimating Models with Sample Selection Bias: A Survey," *Journal of Human Resources* 33 (Winter 1998): 127-169.

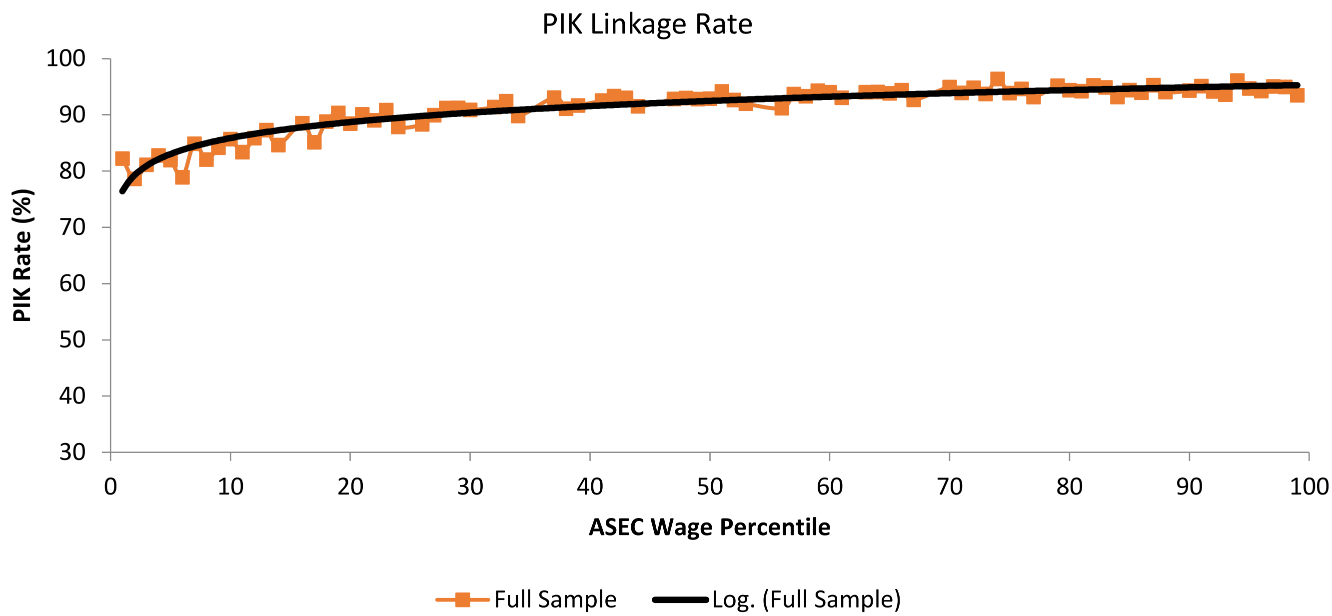
Welniak, Edward J. "Effects of the March Current Population Survey's New Processing System On Estimates of Income and Poverty," Proceedings of the American Statistical Association, 1990.

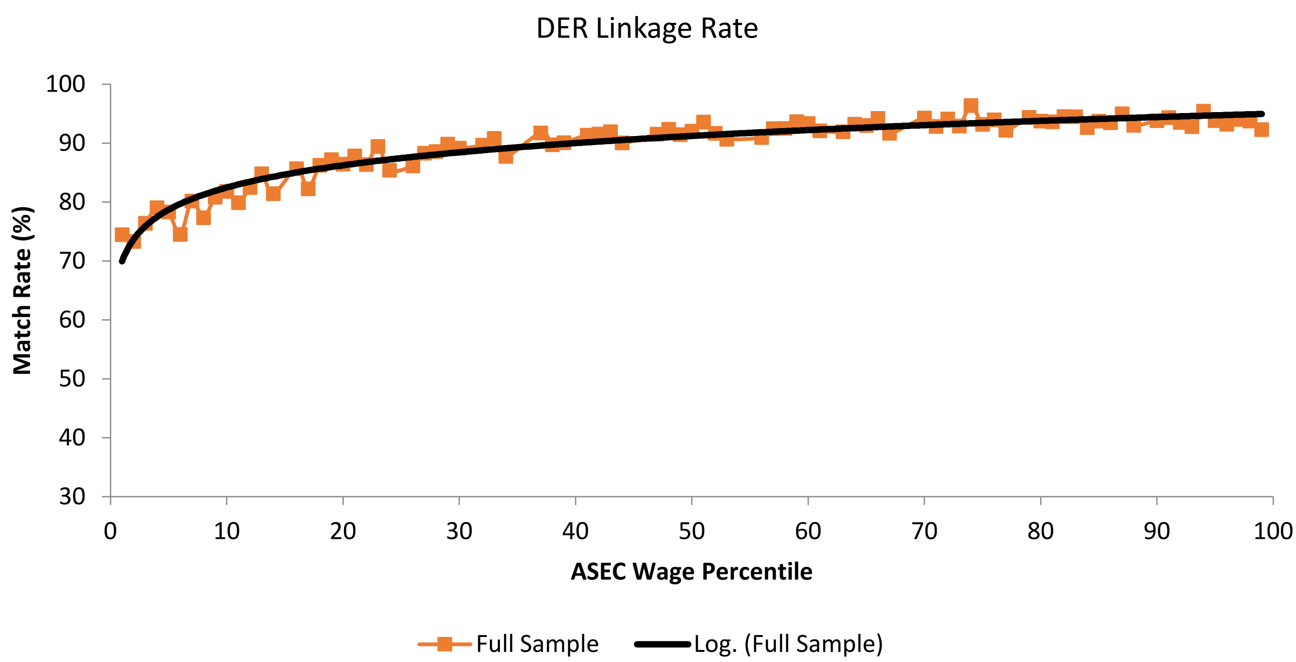
Figure 1: Trends in Item and Total (Item + Supplement) Earnings Imputations in the ASEC



Source: Authors' calculations. U.S. Census Bureau, Current Population Survey, 1988-2012 Annual Social and Economic Supplement.

Figure 2: PIK and DER Linkage Rates across the ASEC Wage Distribution for Combined Male and Female Sample



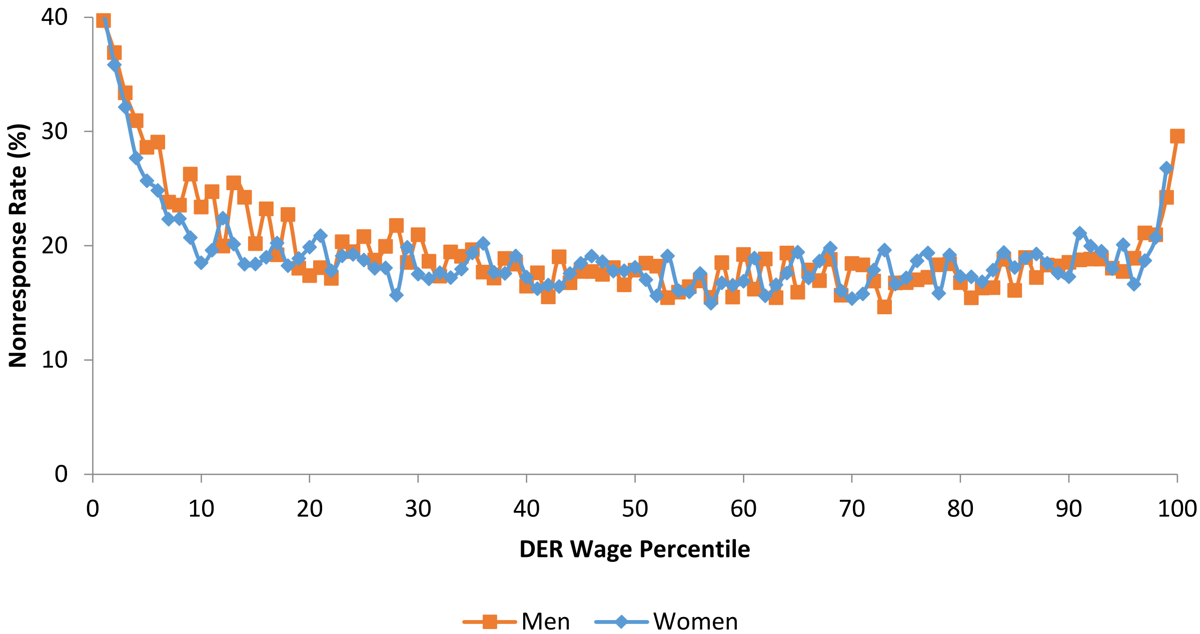


Sources: U.S. Census Bureau, Current Population Survey, 2006-2011 Annual Social and Economic Supplement. Social Security Administration, Detailed Earnings Record, 2005-2010.

Figure 3: Earnings Nonresponse Rates and Conditional Response Rates Relative to Mean by Percentiles over the Male and Female DER Wage Distributions

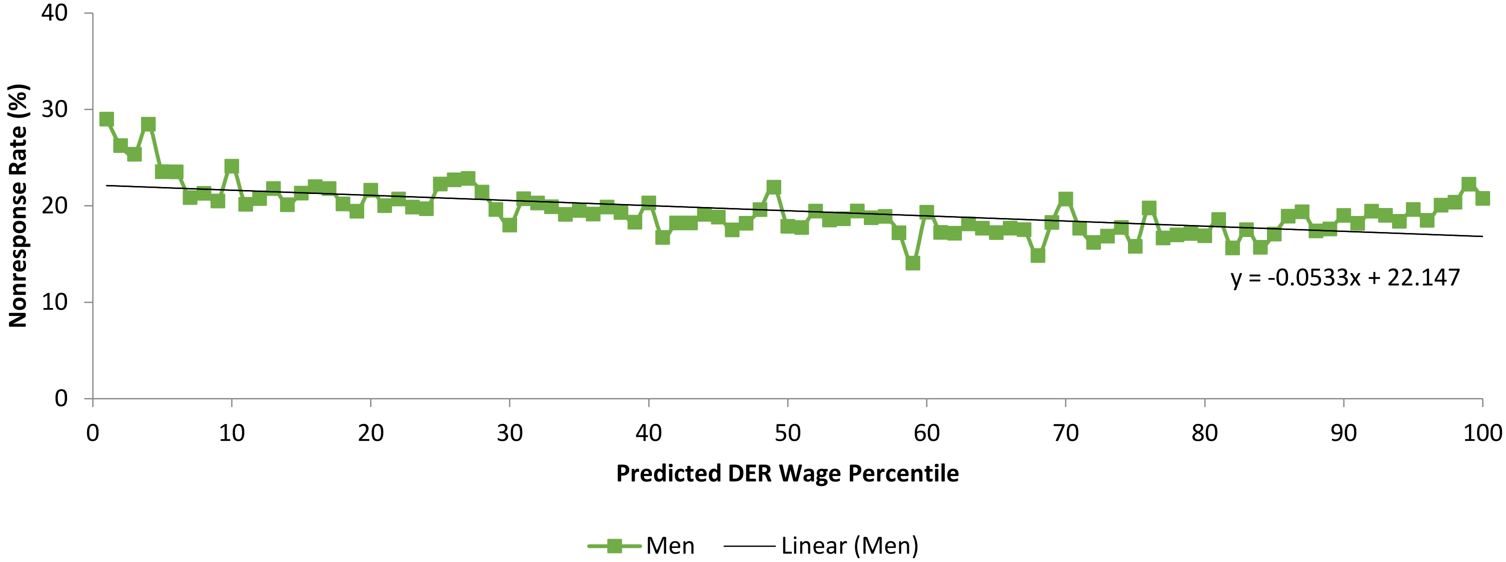
Sources: U.S. Census Bureau, Current Population Survey, 2006-2011 Annual Social and Economic Supplement. Social Security Administration, Detailed Earnings Record, 2005-2010.

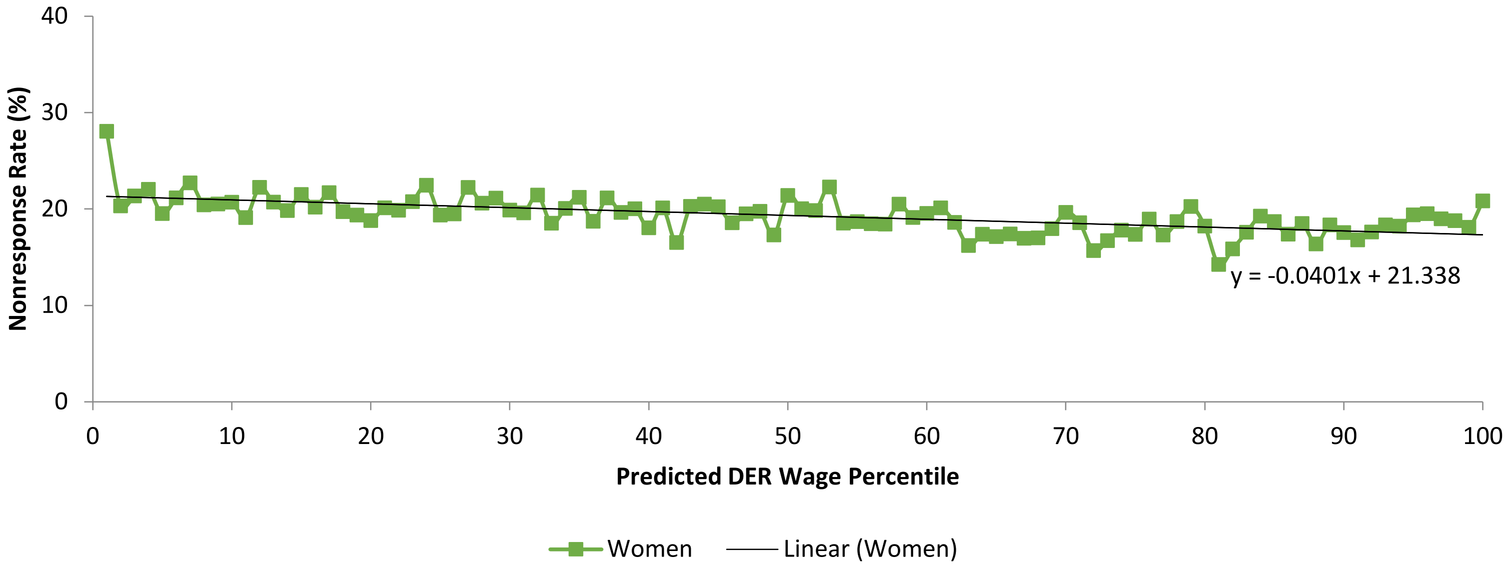
Figure 4: Earnings Nonresponse Rates by Percentile for Men and Women over the Joint Male-Female DER Wage Distribution



Sources: U.S. Census Bureau, Current Population Survey, 2006-2011 Annual Social and Economic Supplement. Social Security Administration, Detailed Earnings Record, 2005-2010.

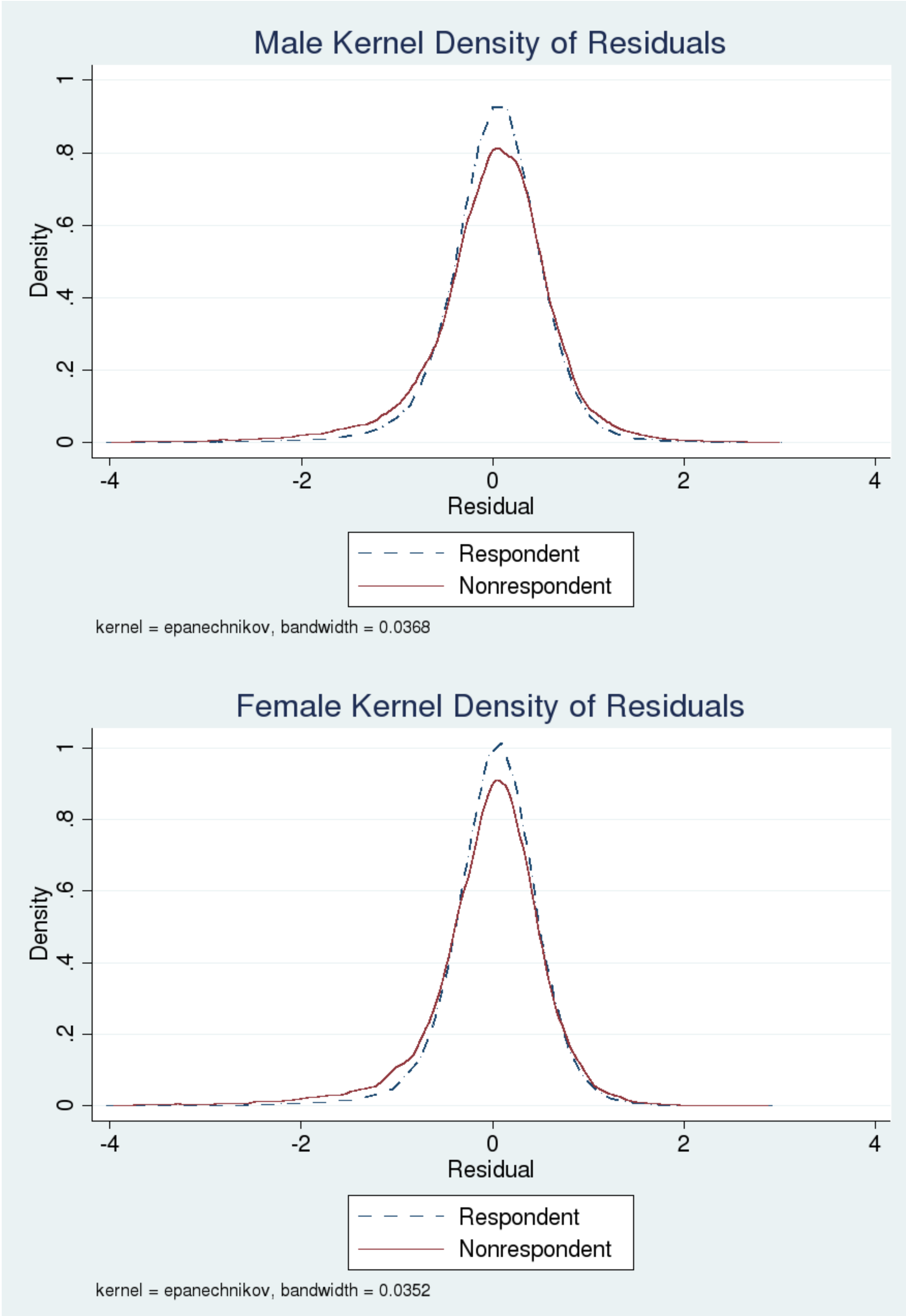
Figure 5: Nonresponse Rates by Predicted DER Wage for Men and Women





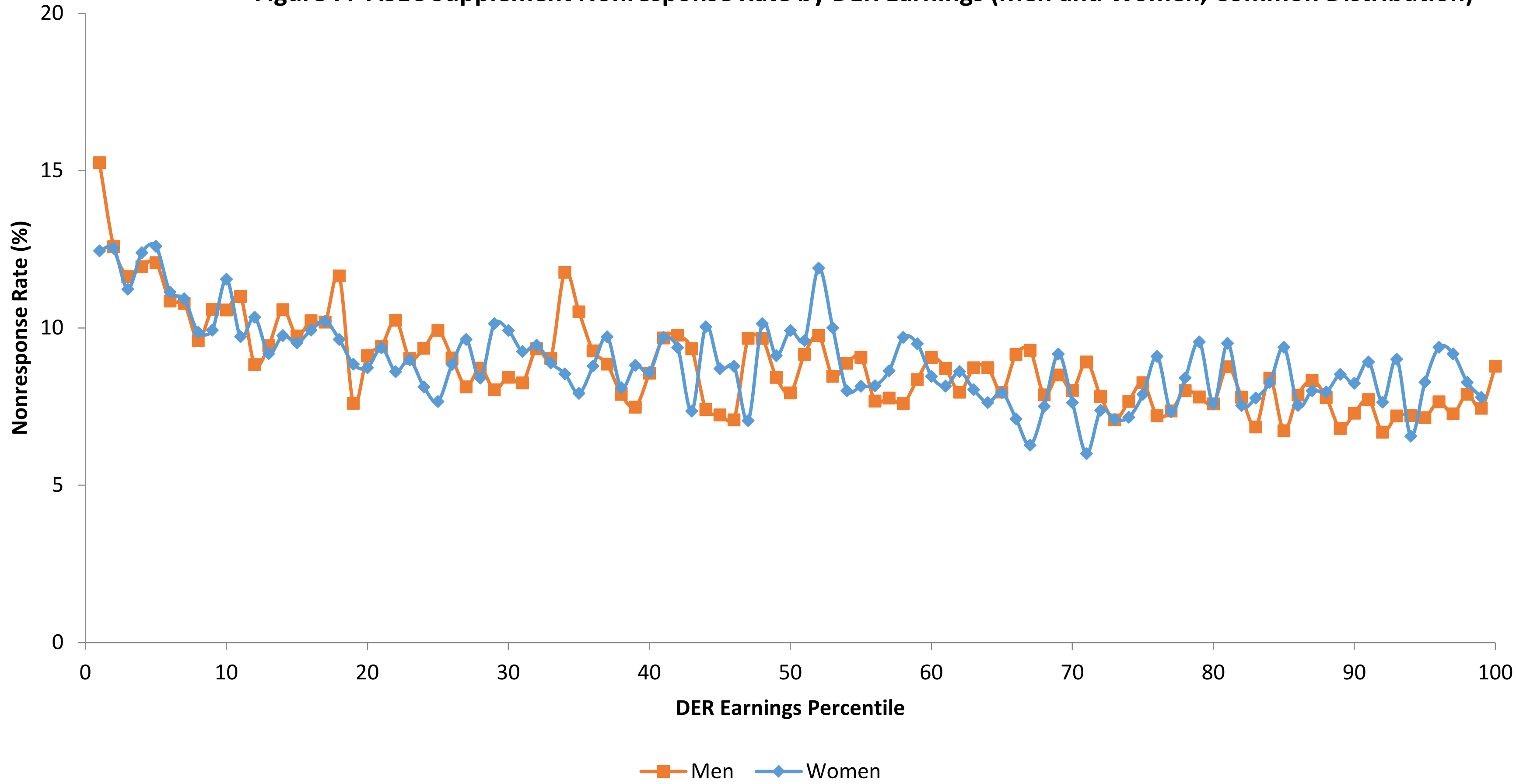
Sources: U.S. Census Bureau, Current Population Survey, 2006-2011 Annual Social and Economic Supplement. Social Security Administration, Detailed Earnings Record, 2005-2010.

Figure 6: Residuals of lnW_i^{DER} Regressions by Response Status in ASEC, by Gender



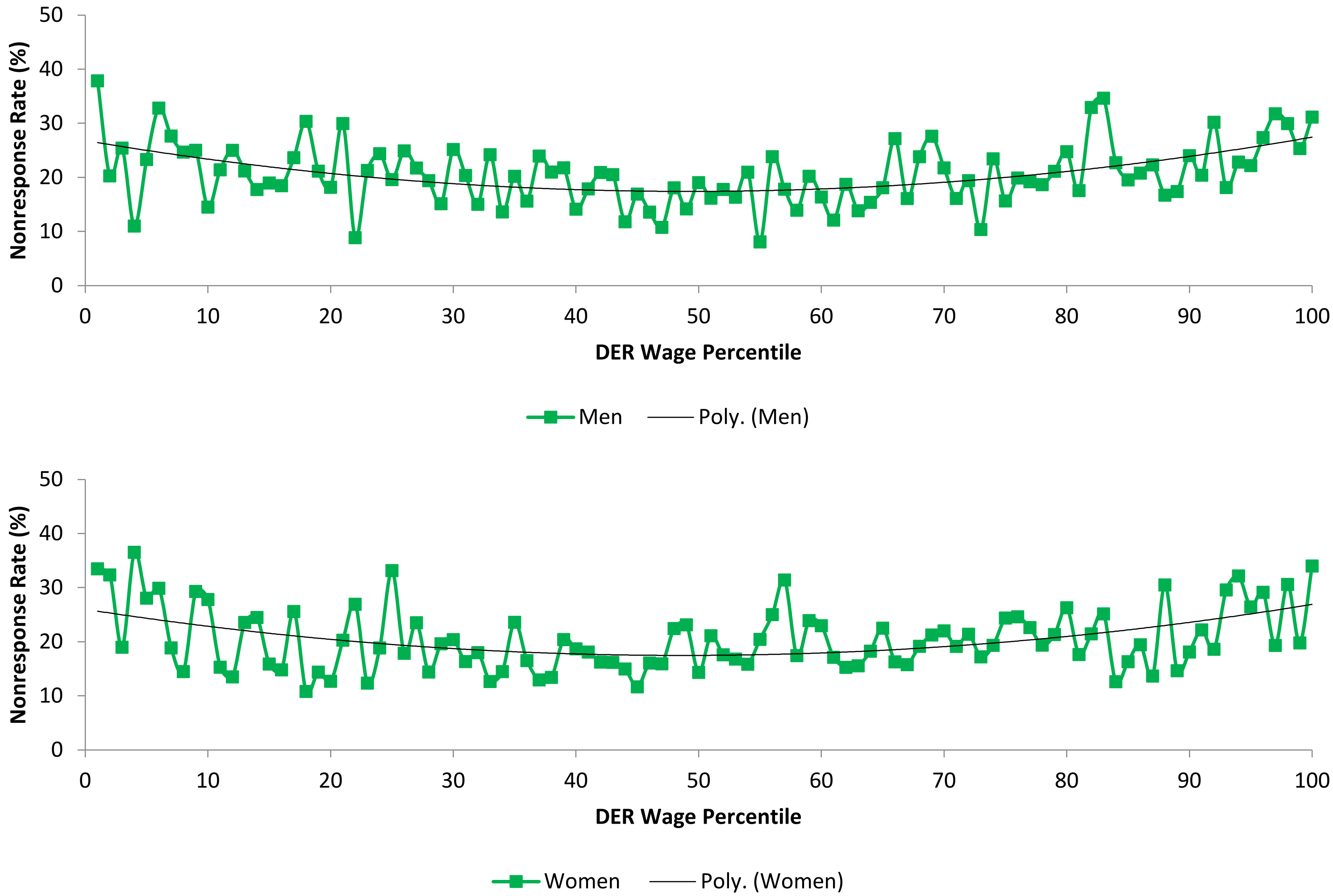
Sources: U.S. Census Bureau, Current Population Survey, 2006-2011 Annual Social and Economic Supplement. Social Security Administration, Detailed Earnings Record, 2005-2010.

Figure 7: ASEC Supplement Nonresponse Rate by DER Earnings (Men and Women, Common Distribution)



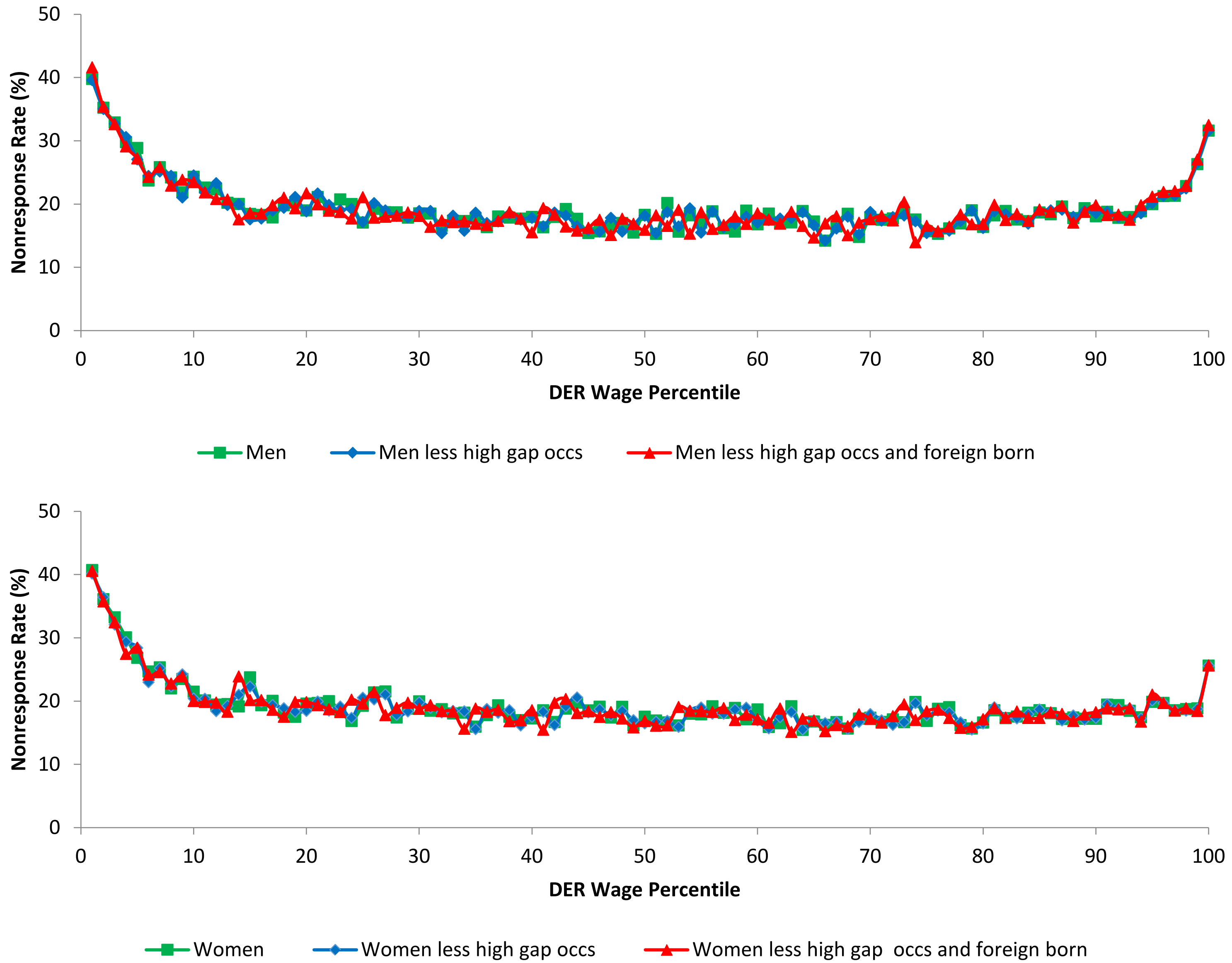
Sources: U.S. Census Bureau, Current Population Survey, 2006-2011 Annual Social and Economic Supplement. Social Security Administration, Detailed Earnings Record, 2005-2010.

Figure 8: Nonresponse Rates by DER Wage for Workers Excluded from Sample, Male and Female Students and Non FT/FY



Sources: U.S. Census Bureau, Current Population Survey, 2006-2011 Annual Social and Economic Supplement. Social Security Administration, Detailed Earnings Record, 2005-2010.

Figure 9: Nonresponse Rates by DER Wage Percentiles for Full Sample and Full Sample minus High Gap Occupations and Foreign Born Noncitizens



Sources: U.S. Census Bureau, Current Population Survey, 2006-2011 Annual Social and Economic Supplement. Social Security Administration, Detailed Earnings Record, 2005-2010.

Table 1: Selected Summary Statistics for Estimation Sample

	Men		Women		
Characteristic	Mean	Std. Dev.	Mean	Std. Dev.	Difference
ASEC Wage ($2010)					
Full Sample	$27.05	27.14	$20.80	18.57	$6.26
ASEC Respondents	$27.11	26.16	$20.94	18.37	$6.18
ASEC Nonrespondents	$26.81	30.83	$20.22	19.38	$6.59
lnW (CPS)	3.075	0.652	2.849	0.604	0.226
DER Wage ($2010)					
Full Sample	$28.24	33.36	$20.80	17.88	7.45
ASEC Respondent	$28.03	30.66	$20.91	16.85	7.11
ASEC Nonrespondent	$29.14	42.71	$20.31	21.65	8.83
lnW (DER)	3.053	0.772	2.818	0.690	0.24
Nonresponse Rate (%)	19.5	39.6	19.3	39.5	0.2
Proxies (%)	53.2	49.9	41.2	49.2	12.0
Observations	157,041		130,663		

Note: All means are weighted using ASEC Supplement weights.

Sources: U.S. Census Bureau, Current Population Survey, 2006-2011 Annual Social and Economic Supplement. Social Security Administration, Detailed Earnings Record, 2005-2010.

Table 2: ASEC Mean Nonresponse with Respect to DER Wages for Men, Women, and 'Mad Men', 2006-2011

	(1) OLS	(2) Probit Marginal Effects	(3) OLS w/X's	(4) Probit w/X's Marginal Effects
		Men		
lnW^{DER}	-0.0268*	-0.0250*	-0.0225*	-0.0200*
	(0.00183)	(0.00167)	(0.00232)	(0.00206)
Constant	0.277*		0.358*	
	(0.00583)		(0.0163)	
Observations	157,041	157,041	157,041	157,041
R-squared	0.003		0.020	
		Women		
lnW^{DER}	-0.0364*	-0.0340*	-0.0418*	-0.0378*
	(0.00211)	(0.00194)	(0.00269)	(0.00235)
Constant	0.295*		0.361*	
	(0.00617)		(0.0179)	
Observations	130,663	130,663	130,663	130,663
R-squared	0.004		0.020	
		Mad Men †		
lnW^{DER}	0.0193*	0.0183*	0.0122*	0.0111*
	(0.00280)	(0.00266)	(0.00347)	(0.00324)
Constant	0.112*		0.203*	
	(0.00926)		(0.0372)	
Observations	78,179	78,179	78,179	78,179
R-squared	0.001		0.018	

* $p<0.01$

† 'Mad Men' sample includes married, white, male U.S. citizens with spouse present.

Estimates are weighted using ASEC Supplement weights.

Sources: U.S. Census Bureau, Current Population Survey, 2006-2011 Annual Social and Economic Supplement. Social Security Administration, Detailed Earnings Record, 2005-2010.

Table 3: ASEC Nonresponse across the DER Wage Distribution for Men, Women, and 'Mad Men', 2006-2011

	(1)	(2)	(3)	(4)	(5)	(6)
	Men		Women		Mad Men†	
DER Wage Deciles and Percentiles	Wage Decile Dummies OLS	Wage Decile Dummies and X's, OLS	Wage Decile Dummies OLS	Wage Decile Dummies and X's, OLS	Wage Decile Dummies OLS	Wage Decile Dummies and X's, OLS
Decile 10	0.288*	0.192*	0.285*	0.176*	0.197*	0.00912
	(0.00434)	(0.0105)	(0.00465)	(0.0114)	(0.00547)	(0.0184)
Decile 20	0.200*	0.111*	0.197*	0.0888*	0.163*	-0.0256
	(0.00411)	(0.0104)	(0.00433)	(0.0113)	(0.00544)	(0.0184)
Decile 30	0.191*	0.106*	0.194*	0.0834*	0.142*	-0.0465
	(0.00389)	(0.0105)	(0.00418)	(0.0113)	(0.00488)	(0.0182)
Decile 40	0.175*	0.0909*	0.178*	0.0678*	0.158*	-0.0315
	(0.00375)	(0.0105)	(0.00403)	(0.0114)	(0.00509)	(0.0183)
Decile 50	0.170*	0.0859*	0.182*	0.0718*	0.169*	-0.0203
	(0.00372)	(0.0106)	(0.00404)	(0.0114)	(0.00529)	(0.0184)
Decile 60	0.173*	0.0908*	0.178*	0.0683*	0.155*	-0.0343
	(0.00372)	(0.0106)	(0.00402)	(0.0115)	(0.00501)	(0.0183)
Decile 70	0.171*	0.0917*	0.167*	0.0586*	0.161*	-0.0312
	(0.00369)	(0.0107)	(0.00390)	(0.0115)	(0.00504)	(0.0184)
Decile 80	0.171*	0.0926*	0.174*	0.0631*	0.171*	-0.0206
	(0.00364)	(0.0107)	(0.00400)	(0.0116)	(0.00515)	(0.0185)
Decile 90	0.184*	0.108*	0.177*	0.0675*	0.186*	-0.00677
	(0.00370)	(0.0109)	(0.00402)	(0.0118)	(0.00523)	(0.0186)
Percentiles 91-95	0.187*	0.113*	0.189*	0.0763*	0.200*	0.00492
	(0.00518)	(0.0116)	(0.00575)	(0.0126)	(0.00759)	(0.0196)
Percentile 96	0.213*	0.141*	0.197*	0.0815*	0.231*	0.0377
	(0.0124)	(0.0163)	(0.0130)	(0.0172)	(0.0180)	(0.0254)
Percentile 97	0.213*	0.142*	0.186*	0.0710*	0.250*	0.0531
	(0.0120)	(0.0159)	(0.0128)	(0.0171)	(0.0182)	(0.0256)
Percentile 98	0.228*	0.157*	0.188*	0.0702*	0.283*	0.0871*
	(0.0125)	(0.0164)	(0.0127)	(0.0171)	(0.0190)	(0.0261)
Percentile 99	0.263*	0.189*	0.189*	0.0716*	0.306*	0.105*
	(0.0130)	(0.0167)	(0.0126)	(0.0170)	(0.0193)	(0.0265)
Percentile 100	0.316*	0.238*	0.256*	0.136*	0.342*	0.139*
	(0.0135)	(0.0172)	(0.0142)	(0.0184)	(0.0196)	(0.0267)
Observations	157,041	157,041	130,663	130,663	78,179	78,179
R-squared	0.201	0.215	0.197	0.210	0.179	0.195

Robust standard errors in parentheses. * $p<0.01$

†"Mad Men" sample includes married, white, male U.S. citizens with spouse present.

Estimates are weighted using ASEC Supplement weights.

Sources: U.S. Census Bureau, Current Population Survey, 2006-2011 Annual Social and Economic Supplement. Social Security Administration, Detailed Earnings Record, 2005-2010.

Table 4: Summary Statistics of Residuals from lnW^{DER} Regressions by Response Status and Gender

	Men				Women			
Statistic	All Men	Non-respondents	Respondents	Difference (NR-R)	All Women	Non-respondents	Respondents	Difference (NR-R)
1%	-1.870	-2.497	-1.656	-0.840	-1.791	-2.452	-1.581	-0.871
5%	-0.863	-1.155	-0.803	-0.353	-0.815	-1.117	-0.748	-0.369
10%	-0.591	-0.743	-0.565	-0.178	-0.551	-0.724	-0.518	-0.207
25%	-0.262	-0.308	-0.253	-0.055	-0.243	-0.314	-0.229	-0.085
50%	0.043	0.041	0.044	-0.003	0.039	0.015	0.043	-0.028
75%	0.334	0.363	0.328	0.036	0.305	0.304	0.306	-0.002
90%	0.613	0.672	0.601	0.071	0.559	0.581	0.556	0.025
95%	0.798	0.886	0.781	0.106	0.726	0.767	0.717	0.050
99%	1.304	1.528	1.248	0.280	1.108	1.204	1.090	0.115
Mean	0.008	-0.023	0.015	-0.038	0.000	-0.063	0.014	-0.077
Std Dev	0.590	0.715	0.559	0.155	0.543	0.667	0.509	0.158
Variance	0.349	0.511	0.313	0.198	0.295	0.445	0.259	0.186
Obs	157,041	28,390	128,651		130,663	24,092	106,571	

Mean, Std Dev, and Variance use ASEC Supplement weights. Percentiles do not use ASEC Supplement weights.

Sources: U.S. Census Bureau, Current Population Survey, 2006-2011 Annual Social and Economic Supplement. Social Security Administration, Detailed Earnings Record, 2005-2010.

Table 5: Occupations with Lowest PIK and DER Linkage Rates

COC code	Occupation Description	Linkage Rates	Number	Count
PIK Linkage Rates:				
4140	Dishwashers	61.8%	239	387
2700	Actors	63.6%	7	11
6420	Painters	64.2%	566	881
6330	Drywall installers	64.3%	305	474
6600	Helpers, construction trades	64.8%	116	179
6510	Roofers	65.6%	223	340
6220	Brick masons	66.4%	235	354
6260	Construction laborers	68.3%	1884	2758
6300	Paving and surfacing	70.0%	28	40
6240	Carpet, floor, tile installers	70.1%	256	365
6050	Misc. agriculture workers	71.0%	1152	1622
7610	Helpers-installation, maintenance, and repair	71.2%	37	52
4130	Dining room and cafeteria attendants, bartender helpers	71.3%	290	407
8310	Pressers, textile, garment, and related	71.4%	110	154
4250	Grounds maintenance workers	72.0%	1325	1841
4020	Cooks	73.0%	2823	3866
8320	Sewing machine operators	73.2%	423	578
8450	Upholsterers	74.3%	75	101
4030	Food preparation workers	74.4%	690	927
DER Linkage Rates among those with and without PIK Linkages:				
6460	Plasterers and stucco masons	47.1%	16	34
4140	Dishwashers	54.5%	211	387
6330	Drywall installers	57.2%	271	474
6420	Painters, construction and maintenance	57.8%	509	881
6600	Helpers, construction trades	59.8%	107	179
6510	Roofers	60.3%	205	340
6240	Carpet, floor, and tile installers	61.4%	224	365
6220	Brickmasons, blockwashers, and stone masons	61.6%	218	354
6050	Misc. agriculture workers	61.8%	1003	1622
6260	Construction laborers	62.0%	1710	2758
6100	Fishers and related fishing workers	63.0%	17	27
2700	Actors	63.6%	7	11
4500	Barbers	64.2%	61	95
8310	Pressers, textile, garment, and related materials	64.3%	99	154
6300	Paving, surfacing, and tamping equipment operators	65.0%	26	40
7610	Helpers--installation, maintenance, repair workers	65.4%	34	52
4130	Dining room and cafeteria attendants and bartender helpers	65.8%	268	407
6250	Cement masons, concrete finishers, and terrazzo workers	65.9%	137	208

Sources: U.S. Census Bureau, Current Population Survey, 2006-2011 Annual Social and Economic Supplement. Social Security Administration, Detailed Earnings Record, 2005-2010.

Table 6: Occupations with Largest Gaps between ASEC and DER Earnings

COC code	Occupation Description	Mean $lnW^{ASEC} - lnW^{DER}$	Count
6460	Plasterers and stucco masons	0.91	16
4500	Barbers	0.80	31
4950	Door-to-door sales workers	0.67	99
4160	Food prep and serving related workers	0.60	7
4920	Real estate brokers and sales agents	0.56	543
6710	Fence erectors	0.55	35
3260	Health diagnosing and testing practitioners	0.52	5
4340	Animal trainers	0.47	24
2700	Actors	0.47	7
4420	Ushers, lobby attendants, ticket takers	0.42	27
2750	Musicians, singers, and related workers	0.40	74
6500	Reinforcing iron and rebar workers	0.39	22
205	Farmers, ranchers, and other agricultural managers	0.39	48
6510	Roofers	0.37	184
4520	Miscellaneous personal appearance workers	0.36	130
4040	Bartenders	0.36	536
3630	Massage therapists	0.35	67
500	Agents and managers of artists, performers, and athletes	0.34	52
8540	Woodworking machine setters, operators, and tenders	0.33	60
6100	Fishers and related fishing workers	0.31	11

Estimates use ASEC Supplement weights.

Sources: U.S. Census Bureau, Current Population Survey, 2006-2011 Annual Social and Economic Supplement. Social Security Administration, Detailed Earnings Record, 2005-2010.

Table 7: lnW^{DER} Wage Equation Predicted Log Earnings with Full Sample, ASEC Respondents, and ASEC Nonrespondents, 2006-2011

VARIABLES	(1) Using betas from lnW^{DER} All Workers	(2) Using betas from lnW^{DER} Respondents	(3) Using betas from lnW^{DER} Non-respondents
Men			
Prediction with full sample X's	3.066	3.075	3.035
Prediction with respondent sample X's	3.074	3.082	3.044
Observations	157,041	128,651	28,390
R-squared of earnings equation	0.384	0.391	0.374
Women			
Prediction with full sample X's	2.809	2.825	2.744
Prediction with respondent sample X's	2.815	2.830	2.750
Observations	130,663	106,571	24,092
R-squared of earnings equation	0.370	0.393	0.311

See text for full discussion. Regression estimates use ASEC Supplement weights. Calculations discussed in text are shown below:

Men:
Bias using respondent sample betas rather than full sample betas with full sample X's:
diff of col 2 and col 1 betas, using DER full sample X's: 3.075 – 3.066 = 0.009
Difference between use of respondent versus non-respondent betas using full sample X's:
diff of col 2 minus col 3 DER full sample X's: 3.075 – 3.035 = 0.040

Women:
Bias using respondent sample betas rather than full sample betas with full sample X's:
diff of col 2 and col 1 betas, using DER full sample X's: 2.825 – 2.809 = 0.016
Difference between use of respondent versus non-respondent betas using full sample X's:
diff of col 2 minus col 3 DER full sample X's: 2.825 – 2.744 = 0.081

Sources: U.S. Census Bureau, Current Population Survey, 2006-2011 Annual Social and Economic Supplement. Social Security Administration, Detailed Earnings Record, 2005-2010.

Table 8: OLS lnW^{DER} Equations, Selected Coefficients for Men and Women, 2006-2011

VARIABLES	(1) All Men	(2) Male Respondents	(3) All Women	(4) Female Respondents
Black	-0.167$^{\Psi}$	-0.155	-0.0865	-0.0869
	(0.00757)	(0.00779)	(0.00573)	(0.00596)
Married, Spouse Present	0.0680	0.0654	0.00111	-0.000975
	(0.00826)	(0.00852)	(0.00612)	(0.00630)
Foreign Born, US Citizen	-0.0764$^{\Psi}$	-0.0693	-0.0386$^{\Psi}$	-0.0493
	(0.00880)	(0.00895)	(0.00846)	(0.00847)
Foreign Born, Not a US Citizen	-0.180$^{\Psi}$	-0.170	-0.163	-0.165
	(0.00985)	(0.0103)	(0.0103)	(0.0107)
GED	0.0738	0.0824	0.148	0.147
	(0.0197)	(0.0207)	(0.0239)	(0.0255)
High School	0.202	0.206	0.261	0.264
	(0.0154)	(0.0165)	(0.0193)	(0.0208)
Some College	0.290	0.293	0.354	0.352
	(0.0158)	(0.0169)	(0.0196)	(0.0210)
Associate's Degree	0.370	0.370	0.466	0.460
	(0.0160)	(0.0172)	(0.0200)	(0.0214)
BA Degree	0.576	0.571	0.649	0.642
	(0.0163)	(0.0173)	(0.0200)	(0.0214)
MA Degree	0.753	0.741	0.842$^{\Psi}$	0.823
	(0.0173)	(0.0184)	(0.0207)	(0.0221)
Professional Degree	1.116$^{\Psi}$	1.068	1.149$^{\Psi}$	1.109
	(0.0239)	(0.0256)	(0.0278)	(0.0285)
PhD	0.963$^{\Psi}$	0.941	1.089$^{\Psi}$	1.069
	(0.0221)	(0.0233)	(0.0255)	(0.0270)
Observations	157,041	128,651	130,663	106,571
R-squared	0.384	0.391	0.370	0.393

$^{\Psi}$ Indicates difference between columns 1 and 2 or 3 and 4 statistically significant at 5% level. Robust standard errors in parentheses. Regressions include quartic potential experience, Asian and other race, separated-divorced-widowed, dummies for 9-12th grade, city size, occupation, industry, region, and year. Estimates are weighted using ASEC Supplement weights.
Sources: U.S. Census Bureau, Current Population Survey, 2006-2011 Annual Social and Economic Supplement. Social Security Administration, Detailed Earnings Record, 2005-2010.

Appendix: The ASEC Imputation Procedure for Earnings

The Census Bureau has used a hot deck procedure for imputing missing income since 1962. This procedure makes the crucial assumption that income data are missing at random. The current system has been in place with few changes since 1989 (Welniak 1990). The ASEC uses a sequential hot deck procedure to address item nonresponse for missing earnings data. The sequential hot deck procedure assigns individuals with missing earnings values that come from individuals ("donors") with similar characteristics. The hot deck procedure for the ASEC earnings variables relies on a sequential match procedure. First, individuals with missing data are divided into one of 12 allocation groups defined by the pattern of nonresponse. Examples include a group that is only missing earnings from longest job or a group that is missing both longest job information and earnings from longest job. Second, an observation in each allocation group is matched to a donor observation with complete data based on a large set of socioeconomic variables, the match variables. If no match is found based on the large set of match variables, then a match variable is dropped and variable definitions are collapsed (i.e., categories are broadened) to be less restrictive. This process of sequentially dropping a variable and collapsing variable definitions is repeated until a match is found. When a match is found, the missing earnings amount is substituted with the reported earnings amount from the first available donor or matched record. The missing earnings amount does not come from an average of the available donors.

The sequential hot deck used in the ASEC is a variant of a cell hot deck procedure, but quite different from the cell hot deck used in the CPS monthly outgoing rotation group earnings files (CPS ORG). Unlike the ASEC procedure, the CPS ORG cell hot deck always requires an exact match on a given set of characteristics with fixed category ranges (i.e., match variables are never eliminated or collapsed). It replaces missing earnings with earnings from the most recent donor having the same set of characteristics. All cells (combinations of attributes) are stocked with a donor, sometimes with donors from previous months. Because all nonrespondents are matched based on the same set of attributes, this makes it relatively straightforward to derive an exact match bias formula (Bollinger and Hirsch 2006) and, more generally, for researchers to know a priori how the inclusion of imputed earners in their analysis is likely to bias statistical results.

The sequential hot deck used in the ASEC has the advantage that it always finds a match within the current month. It has the disadvantage that one cannot readily know which characteristics are matched and the extent to which variable categories have been collapsed. The quality of an earnings match depends on how common are an individual's attributes (Lillard, Smith, and Welch, 1986). Use of a cell hot deck in the ASEC like that used in the CPS ORG would not be feasible. Reasonably detailed matching would require reaching back many years in time to find donors. To insure exact matches within the same month would require that only a few broadly defined match variables could be used, thus lowering the quality of donor matches and imputed earnings.

The ASEC also uses a hot deck procedure for what Census refers to as whole imputes. Whole imputation refers to a household who has participated in the monthly CPS, but refused participation in the ASEC supplement. In this case the entire supplement is replaced (imputed) by a “similar” household that participated in the supplement. The whole imputation procedure uses 8 allocation groups. The set of match variables is smaller than the set used for item nonresponse, consisting of variables available from the monthly CPS for both the supplement nonrespondent and donor household. Like the sequential hot deck procedure for item nonresponse, the match process sequentially drops variables and makes them less restrictive until a donor is found. This requirement implies that donors do not have to answer all the ASEC questions and can have item imputations. Whole imputes account for about 10% of all ASEC supplement records.