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, U.S. Census Bureau

James P. Ziliak, University of Kentucky

June 2017

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 on earnings gaps and inequality using internal Current Population Survey individual records linked to Social Security administrative data on earnings for calendar years 2005-2010. We find that nonresponse across the earnings distribution is U-shaped—left-tail "strugglers" and right-tail "stars" are least likely to report earnings. Throughout most of the earnings distribution nonresponse is ignorable, but there exists trouble in the tails. Approximately half the difference in inequality measures between the CPS and administrative data is accounted for by nonresponse in the CPS. Wage differentials by race, gender, and education in lower and upper quantiles are biased upwards of 20 percent from nonresponse.

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

JEL Codes: J31 (Wage Level and Structure)

We thank Adam Bee, Dan Black, Charlie Brown, James Heckman, Bruce Meyer, Chuck Nelson, Trudi Renwick, James Spletzer, Ed Welniak, and five anonymous reviewers for helpful comments, plus participants at presentations at the U.S. Census Bureau, Society of Labor Economists, Joint Statistical Meetings, American Economic Association Meetings, Brigham Young University, Emory University, Federal Reserve Bank of Cleveland, Federal Reserve Board, Institute for Fiscal Studies, University of Essex, University of New South Wales, and University of Sydney. The analysis provided in this paper has been conducted at the U.S. Census Bureau and in the Atlanta Research Data Center. All results shown in the paper have received clearance from Census. The opinions and conclusions are solely those of the authors.

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. For the Annual Social and Economic Supplement (ASEC, known historically as the March supplement), Census uses a sequential hot deck procedure to address item nonresponse for missing earnings data by assigning individuals with missing earnings values that come from individuals ("donors") with similar characteristics.[1] Less well known is that in addition to item nonresponse, there exists ASEC supplement nonresponse. This occurs when households participating and responding in the monthly CPS refuse to participate in the ASEC supplement. The Census also uses a hot deck procedure for whole supplement nonresponse. Offsetting the progress in data processing, however, were sharply rising rates of earnings and income nonresponse in the 1990s and early 2000s. As depicted in Figure 1, since about 2000 the ASEC has had item nonresponse rates on annual earnings of about 20 percent, while an additional 12 percent of households have the whole ASEC imputed.[2] The rates of item nonresponse are double those at the time of the LSW critique. Additionally, the CPS monthly outgoing rotation group (ORG) files have earnings item nonresponse rates currently well above 30 percent, while the much larger American Community Survey (ACS) has earnings nonresponse rates of about 20 percent, suggesting that nonresponse is pervasive across the most important Federal household surveys.

Unfortunately, we know surprisingly little about patterns of earnings nonresponse, or its potential consequences for important labor-market issues such as earnings gaps by gender and race, or inequality. LSW (1986, p. 492) suggested that ASEC nonresponse is likely to be highest in the tails of the distribution, but provided limited evidence since they could not observe earnings for nonrespondents. LSW (1986, Table 2, p. 493) place white men in eight earnings intervals based on a combination of reported and predicted earnings. They find a U-shaped nonresponse pattern with respect to earnings, with the highest rates in the top three earnings categories.

Whether and to what extent earnings nonresponse is of economic consequence depends on the questions being addressed and the reasons for nonresponse. Prior research has shown that use of imputed earnings can seriously bias average wage gap estimates studied widely by social science researchers (Hirsch and Schumacher 2004; Bollinger and Hirsch 2006, Heckman and LaFontaine 2006) even if the

[1] Welniak (1990) documents changes over time in Census hot deck methods for the March CPS.
[2] An additional form of nonresponse is so-called unit nonresponse, which occurs when there is a noninterview or refusal to participate even in the monthly CPS survey. These rates for the basic CPS were between 8 and 9 percent during our sample period (Dixon 2012). Also, as a point of comparison, nonresponse rates for typical labor supply variables (weeks worked or hours per week) were in the 3 percent - 5 percent range over the past two decades.

earnings data are "missing completely at random" (MCAR). If earnings are MCAR, then nonresponse is not dependent on earnings, even absent covariates; if earnings are "missing at random" (MAR), then nonresponse is not dependent on earnings after conditioning on covariates; and if earnings are "not missing at random" (NMAR), then nonresponse depends on the value of missing earnings even after conditioning on covariates (Rubin 1976; Little and Rubin 2002). It is this latter case that is referred to as having nonresponse bias or nonignorable nonresponse. Both Census imputation procedures and common methods to deal with nonresponse assume that nonresponse is ignorable; that is, those not reporting earnings have earnings similar to those with equivalent measured attributes. If the MAR assumption is violated, measures of earnings gaps and distributions will be biased.

Given the high earnings nonresponse rates in Census household surveys, coupled with a paucity of evidence on nonresponse patterns and its consequences, we address three important and closely-related questions. We do so using restricted-access ASEC household files linked to administrative tax data from the Social Security Detailed Earnings Records (DER) for March 2006-2011 (corresponding to calendar years 2005-2010). First, how does earnings nonresponse vary across the earnings distribution? Access to the DER is uniquely advantageous to address this question as it affords the opportunity to fill in missing earnings for nonrespondents, and to compare survey responses to administrative tax records for respondents. We align each worker's ASEC earnings response status against their corresponding earnings level from the DER. We examine this for men and women separately, as well as for full-time/full-year workers and those whose ASEC reports are provided by a proxy (i.e., another household member). Nearly half of ASEC earnings reports are from proxies. The extent to which proxy reporting affects nonresponse patterns and earnings accuracy are not well understood.

The second question we address, closely related to the first, is whether nonresponse is ignorable. That is, do respondents and nonrespondents have equivalent conditional earnings distributions, and if so, can the earnings of survey respondents accurately describe the missing counterfactual distribution of a combined respondent and nonrespondent sample as if the nonrespondents had responded? This question directly addresses the efficacy of the MAR assumption used in Census imputation procedures, and more broadly, in many related missing-data procedures. MAR relies on the assumption that the joint distribution of earnings and response status, conditional on covariates, can be expressed as the product of the conditional marginal distribution of response status and the conditional marginal distribution of earnings. This leads to our two complementary tests of MAR made possible with access to the DER, one which examines whether the decision to respond to the ASEC earnings question is independent of earnings, conditional on a rich set of demographics, and the second which examines whether the distribution of earnings is independent of response status, again controlling for the same set of demographics. Furthermore, we also estimate parametric and nonparametric earnings regressions using

the DER, and then test 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. Absent the link to the DER these tests are not possible because of missing earnings of nonrespondents in the ASEC.

Third, does earnings nonresponse affect standard estimates of earnings gaps by gender, race, and ethnicity, returns-to-schooling, and earnings inequality and volatility? To address these questions, we estimate saturated quantile earnings models to test how nonresponse affects outcomes in the tails of the distribution, alongside models of central tendency. In addition, we also present estimates of how nonresponse impacts standard measures of inequality such as the Gini, 90-50 and 50-10 ratios, and top income shares. For the volatility analysis, we exploit the longitudinal dimension of the ASEC whereby it is possible to match up to half of the sample from one year to the next to examine both the dynamics of nonresponse as well as implications for summary measures of volatility. Answers to the inequality and volatility 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. Piketty and Saez 2003; Lemieux 2006; Autor et al. 2008; Burkhauser et al. 2012) or fully specified quantile regression models of earnings (Buchinsky 2012; Kline and Santos 2013; Arellano and Bonhomme 2017). Under MAR, unconditioned measures of inequality may differ between the full sample with imputations and a sample omitting imputed earners. And it is not clear a priori which of these two imperfect samples provides the better measure. The full sample is likely to provide a good measure of unconditional inequality if the covariates used in the imputation procedure provide an unbiased measure of earnings and maintain variance. Using only respondents (non-imputes) provides more accurate earnings responses, but risks bias (absent reweighting) to the extent that nonresponse rates differ across the earnings distribution, as we subsequently show. The full sample with imputes is not appropriate for examining conditional inequality, however, because the relationship between inequality and the multivariate correlations with respect to demographic, geographical, and job attributes not used (or used fully) in the imputation process will be biased (Hirsch and Schumacher 2004; Bollinger and Hirsch 2006). Data from the DER are particularly helpful here both because we can fill in the missing ASEC earnings with the DER, and unlike the public-release and internal versions of ASEC, earnings from the DER are not topcoded, which improves our estimates of the importance of nonresponse in the right tail of the earnings distribution.

We are not the first study to examine nonresponse using a validation study, but prior studies are generally old, use small samples, and examine restricted populations (e.g., married white males). Most similar to our initial analysis is a paper by Greenlees et al. (1982), who examined the 1973 ASEC and compared wage and salary earnings the previous year with 1972 linked income tax records of full-time,

full-year male heads of households in the private nonagricultural sector whose spouse did not work. They found evidence that selection into response declined with earnings. David et al. (1986) conducted a related validation study using the 1981 ASEC linked to 1980 IRS reports, also finding evidence of negative selection into response. More recently, Kline and Santos (2013) examined whether returns to schooling and other earnings equation parameters are sensitive to departures from the MAR assumption, using an extract of the 1973 ASEC linked to IRS earnings data, previously analyzed by Bound and Krueger (1991). They provided evidence that missing data probabilities among men are U-shaped, with very low and high wage men least likely to report. Hokayem et al. (2015) used the linked ASEC-DER data to examine how treatment of nonrespondents affects family poverty rate estimates. Although informative and suggestive, it is not known whether results from the earlier studies examining response bias can be generalized outside their time period and narrow demographic samples. In short, there exists little validation evidence regarding the extent and consequences of CPS nonresponse bias across the earnings distribution with recent data. 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.

In general, we find that nonresponse is not ignorable; earnings are not missing at random, 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. This U-shaped pattern is in evidence across gender, race, ethnicity, employment status (hourly and full-time full-year), month-in-sample, proxy earnings status, and panel status (year 1 or year 2).

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 trivial. The exception is for coefficients on variables associated with very high or low earnings. Approximately half the difference in inequality measures between the CPS and administrative data is accounted for by nonresponse in the CPS. Wage differentials by race, gender, and education in lower and upper quantiles are biased by upwards of 20 percent from nonresponse. In short, the economic importance of nonresponse is highly dependent on the research question.

2. Earnings Nonresponse and Response Bias

Official government statistics, as well as most research analyzing earnings (and income)

differences, include both respondents and nonrespondents, replacing the missing earnings with an imputed value. Researchers typically assume (usually implicitly) that nonresponse does not produce systematic biases in the measurement of earnings. The aim of our paper is to determine whether this assumption is justified.

Formally the ignorability of missing earnings (i.e. the MAR assumption) is a statement about the joint distribution of earnings (Y) and response status (R), conditional on covariates (X):

(1) $$f(Y,R|X) = f(Y|X) * f(R|X),$$

which means that once we condition out known demographics and other appropriate covariates, earnings and response status are independent. Because Bayes Theorem permits us to relate the joint distribution of (Y, R) to conditional distributions regardless of whether MAR holds, then we can write the joint distribution as the product of the conditional and marginal distributions:

(2) $$f(Y,R|X) = f(Y|R,X) * f(R|X) = f(R|Y,X) * f(Y|X).$$

The implication of MAR is then readily seen by equating equation (1) and (2)

$$f(Y|R,X) = f(Y|X) \text{ and } f(R|Y,X) = f(R|X).$$

If either of these conditions fails, then the MAR fails to hold. One method for testing for the presence of nonresponse bias across the joint distribution in equation (2) is to treat response as a form of sample selection and to estimate a flexible quantile model via copula methods (Joe 2015; Arellano and Bonhomme 2017). Bollinger and Hirsch (2013) adopted a restrictive version of this approach by estimating the conditional mean of earnings controlling for selection via a standard two-step Heckman (1979) method. This approach is appropriate when the only data available are the survey reports. In this study, we have access to linked administrative earnings for both respondents and nonrespondents, thus circumventing the missing data problem in the ASEC. This permits direct tests of nonresponse bias via validation methods.

Specifically, because the MAR assumption conditions out covariates, it is sufficient to test MAR by focusing on the conditional distributions on the right-hand-side of equation (2). Simply put, does response status depend upon earnings or does the distribution of earnings depend upon response status? For the former we estimate models of the form

(3) $$Pr\left(R_i = 1 \middle| Y_i^{DER}, X_i\right) = F\left(\alpha + \gamma Y_i^{DER} + X_i\beta\right) + u_i,$$

where Y_i^{DER} is administrative earnings reports from the DER described in the next section. Because of the very large number of covariates we restrict these tests to parametric estimators (probit and linear probability) so that a test that $f(R|Y,X) = f(R|X)$ amounts to a test of $\gamma = 0$. We consider specifications

that control for Y_i^{DER} in both logarithmic form as well as flexible percentiles. Greenlees et al. (1982) and David et al. (1986) implemented tests along the lines in equation (3) as it 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 have any predictive power, then earnings and response are not independent, and the MAR assumption fails.

For the test of conditional independence of earnings from response we estimate both parametric and nonparametric models of the form

(4) $$Y_i^{DER} = G(\delta + \theta R_i + X_i\pi) + v_i.$$

Summary measures of $f(Y|R, X)$ are the key 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 (1979) and later papers (for a survey, see Vella, 1998) suggest that a key parameter is $E[Y|R = 1, X]$ in which case the test that $f(Y|R, X) = f(Y|X)$ amounts to a test of $\theta = 0$. 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. For the nonparametric models we estimate kernel density functions separately for respondents and nonrespondents and conduct Kolmogorov-Smirnov tests of the null that $f(Y|R = 1, X) = f(Y|R = 0, X)$. Rejecting the null of equality is a sufficient condition to reject the hypothesis that $f(Y|R, X) = f(Y|X)$.

3. Data: The ASEC-DER 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).[3] In addition to the data included in public-use ASEC, the internal ASEC file has values for income sources that are substantially higher than the public use topcodes.[4]

A. Estimation Sample

The DER file is an extract of the Master Earnings File and includes data on total earnings as reported on a worker's W-2 form, wages and salaries and income from self-employment subject to Federal Insurance Contributions Act and/or Self-Employment Contributions Act taxation, as well as deferred wage (tax) contributions to 401(k), 403(b), 408(k), 457(b), and 501(c) retirement and trust plans,

[3] The linked ASEC-DER were obtained as part of an internal-to-Census project and analyzed in a secure facility at the U.S. Census Bureau in Suitland, MD. Researchers outside of Census interested in accessing such data must have their project approved by Census and the Internal Revenue Service for analysis conducted in a secure Federal Statistical Research Data Center. For more information see https://www.census.gov/fsrdc .

[4] Larrimore et al. (2008) document the differences in topcode values between the internal and public use CPS files.

all of which we include in our earnings measure. 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). Unlike the internal ASEC earnings records, the DER earnings are not topcoded.[5] 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. 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.

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"), total self-employment earnings, and total deferred contributions across all employers. In this way, DER earnings are most compatible with ASEC earnings from all wage and salary jobs (WSAL-VAL) plus non-negative self-employment earnings. We classify a worker as having imputed earnings if wage and salary income from the longest job (I-ERNVAL), from other jobs (I-WSVAL), or from self-employment earnings is imputed. For much of our analysis, we focus on annual earnings because earnings are available in both the ASEC and DER, but we also examine earnings among full-time full-year workers, as well as average hourly earnings found 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) from the ASEC and are available earnings item nonrespondents.

Combining wage and salary along with self-employment earnings is particularly advantageous using the linked household and administrative data because earnings are often reported differently in household surveys than in tax records. For example, in occupations such as the clergy and real estate agents, much of their earnings are reported as self-employment earnings in the tax records, but as wage and salary earnings in the ASEC. For such workers, ASEC and DER reports of either wage and salary earnings or self-employment earnings differ substantially between ASEC and DER, whereas combined earnings are far more similar.

The principal sample used in our analysis includes civilian wage and salary workers ages 18 to 65 who have reported or imputed positive earnings in the prior year. We exclude workers who are full time students and workers identified in ASEC who have been linked to the DER but show zero earnings in the

[5] 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.

DER and have positive deferred compensation. We exclude individuals with whole imputes of the ASEC (those for whom all ASEC data are imputed), and provide a separate analysis of this subsample in a robustness section below. The full sample, including those with no ASEC-DER link, consists of 479,043 earners (251,498 men and 227, 545 women).

[Table 1 here]

Table 1 provides summary statistics for our full sample in the first column, weighted by the ASEC person supplement weight. The average worker is 41 years old, slightly more likely to be a male (52.5 percent) and has an average of 13 years of education. The majority are married with spouse present (57 percent), native born (84 percent), and work full time, full year (72 percent). Nonresponse to either the wage and salary questions (longest job and all jobs in prior year) or the self-employment earnings question totals 21.8 percent of the sample. The majority of this is concentrated on the wage and salary questions (21.6 percent) largely because relatively few individuals are self-employed. The structure of the ASEC interviews identifies a single respondent in the household who provides information about other members of the household; hence, 48 percent of the responses are proxy responses, an issue that we return to in a subsequent section. Inflation-adjusted ASEC total earnings is $46,178, while average DER earnings are a higher $48,360, and in both cases the majority of earnings are from wages and salaries.

B. Sample Differences by Response and Linkage Status

Table 1 also presents descriptive statistics for the sample broken down by ASEC response status and by DER link status. The link between the ASEC and DER is done within the Census Bureau's Center for Administrative Records Research and Applications using a unique Protected Identification Key (PIK). The PIK is a confidentiality-protected version of the Social Security Number (SSN). 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.[6] 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 in order to link the ASEC and DER. Figure 2 shows the linkage rates across the ASEC earnings distribution for our full sample of all wage and salary workers, for full-time full-year workers, and by gender. Linkage rates between the ASEC and DER administrative data average about 87 percent. The linkage rate is lowest at the low end of the earnings distribution, with men notably lagging women in linkage rates in the bottom

[6] The Census Bureau changed its consent protocol to link respondents to administrative data beginning 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 link to administrative data; i.e., an "opt-in" consent option. Beginning with the 2006 ASEC, respondents not wanting to be linked to administrative data had to notify the Census Bureau through the survey field representative, website or use a mail-in response in order to "opt-out". This opt-out rate is a very small 0.5 percent of the ASEC sample. If the respondent doesn't opt out, they are 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 only 61 percent, as compared to 87 percent for our full 2006-2011 ASEC sample (Table 1).

quarter of earnings. Below we discuss how we handle missing linkages.

[Figure 2 here]

In general, Table 1 shows that nonrespondents are not markedly different than the full sample (or the respondent sample). They are slightly less likely to be a Hispanic (1 percentage point difference) or a female (2 point difference), more likely to be never married (3.9 point difference), and more likely to be full-time full-year workers (2.5 point difference). However, there are few differences in earnings. In both the ASEC and the DER measures, nonrespondents have slightly higher annual earnings. It is unsurprising that the ASEC difference is small since the imputed earnings derive from the earnings of the responders. The small differences in sample structure have little impact on overall averages.

The nonlinked sample shows more striking differences with the full sample. Individuals for whom a link was not found are nearly two years younger, 7.1 percentage points more likely to be male, and have 1.1 fewer years of education. Most notably they are more than twice as likely to be Hispanic (34.2 percent compared to 14.4 percent), and over three times more likely to be foreign born and not a citizen (30.7 percent compared to 9.1 percent). They are also less likely to be full-time full-year workers, which helps explain their significantly lower earnings reported (or imputed) in the ASEC ($35,552) compared to the full-sample ASEC ($46,178).

[Table 2 here]

Table 2 further divides the full sample into four groups: linked respondents, linked nonrespondents, nonlinked respondents and nonlinked nonrespondents. Linked responders have the highest percentage of women, the highest educational attainment, the highest percent who are married spouse present, and the highest rate of native born. Linked nonrespondents have the highest concentration of full-time full-year workers, and the highest concentration of Blacks. Nonlinked respondents have the highest concentration of Hispanics and males, and the highest percent of part time full year workers. Nonlinked nonrespondents have the highest concentration of single-never married individuals, foreign born citizens, and full-time part-year workers.

Because of possible differences between those linked with a PIK and those not linked (for a review of the CPS linkage, see Bond et al., 2013), in Table 3 we further subdivide the sample based on nativity and Hispanic ethnicity. There we see that Hispanic workers have a lower link rate in the full sample than non-Hispanic workers (other demographic groups did not show marked differences). Among those who are native born, the difference between Hispanic and non-Hispanic samples is not remarkable (87.9 percent as compared to 91.1 percent). Further, among the immigrant Hispanic samples, those who have become naturalized citizens – while exhibiting lower link rates than the native born – the difference is again small (83 percent vs 85.8 percent). However, among those who have not become naturalized

citizens, the difference is quite substantial: only 44.8 percent of non-naturalized Hispanic immigrants in the ASEC were linked to a DER record. Non-naturalized Hispanic immigrants are approximately 6 percent of the full sample, yet account for 26 percent of the nonlink cases. On the other hand, Table 3 shows that the nonresponse rates for the earnings questions are very stable across these groups. Overall 21.8 percent of the ASEC sample failed to respond to the wage and salary or self-employment earnings questions, and this rate is little different among the native born and naturalized and non-naturalized Hispanics (again, nearly all is from wage and salary nonresponse). We note that non-citizen immigrants are dominated by Hispanics at 64.6 percent; naturalized immigrants are 38.9 percent Hispanic.

[Table 3 here]

The evidence in Table 3 suggests that link failure between ASEC and DER is concentrated on undocumented immigrants. Because the opt-out rate to agree to link ASEC and the DER is a trivial 0.5 percent, the majority of the "link failures" are due to a lack of personally identifiable information needed to construct a SSN, and in turn a PIK. In order to address this, we have estimated a saturated probit model of the probability of an ASEC-DER link as a function of a full array of demographic characteristics, including nativity, Hispanic ethnicity, and their interaction (see Appendix Table 1). We then use the fitted values from the probit to construct inverse probability weights (IPW) to rebalance the ASEC-DER linked sample for the missing nonlink sample.[7] Because most of the linkage failures are not due to an opt-out choice by the respondent, and instead are accounted for by observed demographics, we believe any potential bias from selection on unobservables, which would not be corrected by IPW, is minimal.

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

We begin our analysis examining the conditional distribution of response given earnings, where in Table 4 we present estimates of equation (3) using the linked ASEC-DER sample and both IPW linear probability (columns (1) and (3)) and IPW probit (columns (2) and (4)). In this first test, we control for DER earnings simply in logarithmic form. The first two columns do not control for any confounders, while in columns (3) and (4) we control for a rich a set of covariates in X_i, including 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.

[Table 4 here]

[7] Specifically, we deflate the Census ASEC weight by the probability of an ASEC-DER link. In practice, a comparison of the IPW estimates with non-IPW estimates shows minimal differences.

The results in Table 4 suggest a central tendency of positive rather than negative selection into response. That said, the coefficients for both men and women are very close to zero (with or without controls). The marginal effect for men with controls is a precisely estimated -0.01 (a 10 percent increase in earnings decreases the probability of nonresponse by a tenth of a percentage point). The effect for women is half that size (-0.005). 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 less than 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 estimates with those of Greenlees et al., in results not shown 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. In contrast to the negative coefficients on log earnings for all men, using the restrictive married while-male sample flips the signs and produces positive coefficients, meaning negative selection into response. 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 their sample is restrictive and not representative of the modern labor force.

[Figure 3 here]

Rather than focusing on central tendency, it is more informative to examine how nonresponse varies across the distribution. Grouping observations by DER earnings centile for the linked sample and estimating nonresponse rates for each centile (by gender) produces nonresponse rates which vary across the distribution non-parametrically. Panel A of Figure 3 plots these results for the entire sample. We note that the highest nonresponse rates are for men through the lowest 30 centiles (as high as 25 percent) and for both men and women at the highest 5 centiles, reaching 30 percent for men and over 25 percent for women. Throughout the middle of the distribution, the graph is relatively flat. This is suggestive of our main result—"trouble is in the tails"—which is underscored in more dramatic fashion in Panels B and C of Figure 3. In Panel B we focus upon earnings among full-time, full-year workers (72.7 percent of the linked sample), and in Panel C we adjust for hours of work regardless of work status and depict nonresponse rates across the distribution of average real hourly earnings. Here the "trouble in the tails" is most evident: nonresponse rates rise dramatically in the left and right tails. Men and women in the lowest 5 percent of the earnings distribution have nonresponse rates in excess of 30 percent, rising as high as 40 percent in the lowest centile. Similar to Panel A, both men and women in the highest centiles have

nonresponse rates reaching 30 percent. Through the middle of the distribution, however, the nonresponse rates are remarkably flat. The linear models reported in Table 4 will necessarily fit this part of the distribution, thus explaining the apparent absence of substantive nonresponse bias when focusing on central tendency. The less pronounced "trouble" in the lower tails in the top panel – which includes part-time and part-year workers – is largely explained by the fact that low earnings is caused not only by low pay, but also by few weeks worked and low weekly hours. Both Panels B and C adjust for these hours differences, and yield a more striking pattern of U-shaped nonresponse across the distribution.[8] In short, nonresponse is associated primarily with a low wage; not low earnings resulting from low hours worked.

While the level of nonresponse between men and women differs, the general shape of the relationship with earnings percentiles is similar across gender. We turn next to examining nonresponse across earnings for racial and ethnic groups. We focus on four groups here: Non-Hispanic White, Non-Hispanic Black, Hispanic, and Non-Hispanic Asian (includes Native Americans, Pacific Islanders, and other groups). Figure 4 is comparable to Figure 3 in that a U-shaped relationship is muted in Panel A (annual earnings) but readily apparent in both Panels B and C (FT/FY and hourly earnings, respectively). The U-shape is ubiquitous: it appears to hold for all racial and Hispanic ethnic groups. The overall level of nonresponse differs somewhat across groups, with Blacks having the highest rates throughout much of the distribution, and Hispanics having the lowest rates. Differences between Whites, Hispanics, and Asians are very small.

[Figure 4 here]

We also examine nonresponse by interview month in Figure 5. The sample design of the ASEC is such that respondents are in for four months, out for another eight months, and then in for four more months. Interviews for months 1 and 5 are conducted in-person, while months 2-4 and 6-8 typically rely on telephone interviews. Bollinger and Hirsch (2013) note that nonresponse for the ASEC earnings question is lower in the first and fifth month in sample (MIS 1 and 5). Krueger, Mas, and Niu (2017) find rotation group bias in unemployment rates, with the highest rates in MIS 1 and 5. Hirsch and Winters (2016) find the same pattern for multiple job holding. It is not surprising that earnings nonresponse is lower in MIS 1 and 5, but this raises the question of whether the U-shaped nonresponse pattern found for the full sample differs between MIS 1 and 5 and the other months in sample. Figure 5 graphs the average nonresponse rates from months 1 and 5 against the average nonresponse rates in months 2-4 and 6-8. Panels A-C are for men and refer to earnings among all workers, earnings among full-time full-year

[8] Reported hours are concentrated at 2080 (full-time, full-year). While nonresponse is somewhat higher for workers at 2080 hours (3 percentage points), there is no other obvious pattern across hours worked. Mean annual hours worked systematically increase across the DER earnings distribution, as expected. That said, for those with low DER earnings, mean hours worked are substantial, about 1000 hours for men and 650 for women in the lowest 3 earnings percentiles. This suggests that hours worked are not driving the U-shape.

workers, and average hourly earnings of all workers. Panels D-F present parallel figures for women. The graphs highlight that the months spanning telephone interviews have higher nonresponse, but in all cases the U-shape does not depend upon the month in sample.

[Figure 5 here]

The nonresponse figures seen above do not control for other factors, many of which are known to be associated with earnings and nonresponse. To address this, we modify the nonresponse equation specification seen previously in Table 4 by grouping the bottom 90 percent of earners into wage deciles, while breaking up the top decile into finer percentile increments. Table 5 presents results both with and without human capital, demographic, and location controls, separately for men and women. In all cases, we include a full set of decile/percentile dummy variables, rather than including an intercept. Hence, each decile/percentile coefficient provides an estimate of the nonresponse rate at the given DER wage level. Readily evident from the coefficients is that nonresponse rates are not constant across the distribution. Similar to Figure 3, columns (1) and (3) in Table 5 demonstrate a U-shape, with the highest earnings deciles producing the highest nonresponse. This U-shape is even more pronounced in columns (2) and (4), especially for men, where we include a full set of controls similar to Table 4. Among men, the lowest decile has a 14.3 percent nonresponse rate, while the typical range through the rest of the distribution is roughly half that at 7.5 percent to 8.5 percent. For men in the highest 3 percentiles, the nonresponse rate again rises over 14 percent with the top 1 percent having a 20 percent nonresponse rate. For women, the results with controls are less pronounced, but again we see the U-shape. At the lowest decile, the nonresponse rate is 11 percent, while through the middle of the distribution it falls to around 9 percent, and in the highest percentile, it rises to 15 percent. While we do not reject the null hypothesis that these rates are equal through the middle of the decile range (40th through 70th deciles), we do reject the null that all deciles are equal.

[Table 5 here]

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 6. We do this in order to test whether the U-shape is largely a result of observable covariates, or unobservables. The linked ASEC-DER sample was used to estimate conditional mean earnings equations along the lines of equation (4) using a rich set of controls, including education, demographics, location, and job type, as well as controls for both full time/part time and full year/part year status. The predicted DER wage for each worker, which can be thought of as an "attribute index", is then used similarly to the actual DER wage in Figure 3. Workers are grouped by centile and the resulting nonresponse rate is plotted, along with a smoothed quadratic trend function. Panel A of Figure 6 makes it clear that nonresponse is somewhat higher in the tails of the

attribute distribution of men compared to women in Panel B. For the most part, though, nonresponse for men and women demonstrates less of a U-shape across the attribute distribution than it does across the earnings distribution. The U-shaped nonresponse (i.e., trouble in the tails) observed in Figures 3-5 is not driven primarily by observable earnings attributes; rather, it results from the *realization* of either very low or very high earnings.[9]

[Figure 6 here]

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 through the center 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 middle of the distribution, it is one of nonresponse declining ever so slightly with respect to earnings over the distribution before turning up at the very top percentiles. For men, we regard any such pattern between the 30th and 95th percentiles as inconsequential. Where there most clearly exist problems is in the tails. Stated simply, nonresponse is highest among left-tail "strugglers" and right-tail "stars". Characterizing selection into response based solely on estimates of central tendency over entire distributions, as seen in Table 4 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 topcoded in public use ASEC files.[10]

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

[9] Note that our unconditioned figures showing nonresponse rates across the earnings distribution (Figures 3-5) also have been constructed conditioned on a detailed set of covariates. While conditioning affects the level of nonresponse, curvature of the conditioned and unconditioned nonresponse figures are indistinguishable to the eye.
[10] 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 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. More recently, Census has revised their methods and now use a "rank proximity swap" procedure that swaps earnings values among workers reporting high earnings within given bands.

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, and thus lower DER earnings, may also be less willing to answer ASEC earnings questions.

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

In this section, we examine the distribution of earnings conditional on response and earnings covariates, f(Y|R,X), again using the linked ASEC-DER data with inverse probability weights. We estimate wage regressions specified in equation (4) using lnW_i^{DER} and provide kernel density estimates of residuals for respondents and nonrespondents. We also provide summary statistics of the residuals by response status and test for differences between the two distributions.

[Figure 7 here]

The estimated distributions of residuals are presented in Figure 7. The top panel of Figure 7 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 6 supports this, demonstrating that the variance for male (female) nonrespondents is 1.31 (1.08) 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 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 7. 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 (p-value < 0.00) that $f(Y|R=1,X) = f(Y|R=0,X)$, which is a sufficient condition to reject the hypothesis that $f(Y|R,X) = f(Y|X)$.

[Table 6 here]

Information in Table 6 provides an alternative way to present 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 Considerations for Earnings Nonresponse: Proxies, Measurement Error, Whole Imputations, and Earnings Growth

The evidence presented points strongly to the presence of nonrandom nonresponse of earnings in the tails of the distribution. We next provide a complementary set of analyses to examine, in turn, (a) the role of proxy respondents in earnings nonresponse; (b) the potential confounding influence of measurement differences between the ASEC and DER in assessing earnings nonresponse; (c) DER earnings among households who did not participate in the ASEC supplement (whole imputations); and (d) measurement of earnings growth using the panel dimension of the ASEC across years.

A. Proxy Respondents

Census interviewers designate a single person to be the respondent for all household members. Designated respondents are more likely to be women than men. Although a single person is recorded as providing answers to survey questions, in practice the designated respondent may rely on input from other household members in providing requested information. In the ASEC sample used in our analysis, 53.6 percent of men have their earnings reported by a proxy, while 40.5 percent of women rely on proxy reports.[11] This high level of proxy reporting follows automatically from Census survey procedures.

[Figure 8 here]

Earnings nonresponse is substantially higher among individuals with proxy earnings responses than among self-respondents. For our combined sample of women and men, earnings nonresponse rates are 22.8 percent for proxy respondents versus 14.6 percent for self-respondents. The gap in nonresponse rates between proxies and self-respondents is about 5 percentage points greater among men than among women. Earning nonresponse rates by proxy status are shown for men and women in Figure 8 across both

[11] We identify a proxy response by when the line number of the respondent in the household is not equal to an individual's line number. This method is not 100 percent accurate. If there is a change in the respondent after the survey collects earnings information, this comparison will not reflect who the respondent was for the earnings information that was collected. It reflects the respondent when the interview ends.

the DER earnings distribution (Panels A and C) and the average hourly wage distribution (Panels B and D), separately for self-respondents and those whose earnings are reported by a household proxy. Despite the large differences in levels of nonresponse between proxies and self-respondents, their nonresponse patterns are similarly U-shaped as seen previously in Figure 3 using the combined sample. We see high rates of nonresponse in both tails of the distribution, with relatively flat rates throughout the middle of the distribution. A finding that use of proxies increases item nonresponse of earnings and lessens the accuracy of reported earnings need not imply that Census should increase use of self-reports and rely less on proxy respondents. Use of proxies substantially lowers the time and money costs of conducting Census household surveys. Unknown is the extent to which nonresponse among proxies is due to poor information on earnings. For some (unknown) share of proxies, earnings nonresponse may be preferable to a poor guestimate of earnings.

There exists rather limited information on the reliability of proxy earnings responses. Prior studies include Mellow and Sider (1983), who conducted a validation study comparing responses in the January 1977 CPS (about 4,500 workers) with matched data from these workers' employers. They found the difference between log wages reported by the employer minus log wages in the CPS to be an average 0.048, with self-respondents having an average 0.045 and proxy respondents 0.052. Their qualitative finding that proxy respondents understate earnings relative to self-respondents is consistent with our results, presented subsequently. Mellow and Sider do not show separate results by gender. In the analysis that follows, we find that men have earnings understated by proxy respondents (who are often women) more so than do women (who often have male proxies).

Two relatively recent studies examining proxies using public use data (Reynolds and Wenger 2012; Lee and Lee 2012) emphasize gender differences in proxy reporting. Reynolds and Wenger use matched two-year CPS panels and compare wages for individuals that are self-reported versus proxy reported. They find that that individual wage changes (one year apart) for Proxy-to-Self wage reports systematically exceed wage changes for Self-to-Proxy wage reports. Reynolds and Wenger note the relative change over time for women versus men in the use of proxies and highlights how this can lead to mismeasurement of changes in the gender wage gap over time. Lee and Lee (2012) analyze data from the Michigan Panel Study of Income Dynamics (PSID). They highlight the finding that in the PSID an increasing share of household respondents are women, which accounts for some of the (apparent) decline in the gender gap over time.

Using the linked ASEC-DER sample, we can observe whether administrative earnings in DER, where there are no proxies, vary with respect to proxy use in ASEC. That is, we estimate two equations, each separately for men and women: (1) an ASEC wage equation with spouse and nonspouse proxy variables and (2) a DER wage equation with ASEC spouse and nonspouse proxy variables. Each wage

regression also controls for a saturated set of confounders. The proxy variables in the DER equation act as "phantom" dummies; if ASEC proxy coefficients only measured true reporting differences between proxies and self-respondents, the DER proxy coefficients should be zero. Proxy coefficients in ASEC wage equations reflect the combined effects of proxy misreporting and worker heterogeneity. Inclusion of "phantom" ASEC proxy variables in DER administrative earnings regressions (where there are no proxies) provides estimates of worker heterogeneity correlated with proxy status. Subtracting the DER proxy coefficients measuring heterogeneity from the ASEC proxy coefficients produces estimates of proxy misreporting error. The ASEC wage equations exclude imputed earners since we cannot know whether the donor's earnings were self-reported or from a proxy. That same sample is included in the DER wage equation.

[Table 7 here]

In order to estimate proxy misreporting error, we simply subtract the DER phantom proxy coefficients (which reflect unobserved heterogeneity) from the corresponding ASEC proxy coefficients (reflecting unobserved heterogeneity and misreporting). These results are summarized in the two far right columns in Table 7, using both a single proxy variable and distinguishing between spouse and nonspouse proxies. Based on use of a single proxy (i.e., proxy use versus self-response), we find that proxies understate men's annual earnings by a substantive 5.9 log points and hourly earnings by 5.5 log points. Underreporting by wives and nonspouse proxies is nearly identical. The DER results differ substantially with the ASEC coefficients, where nonspouse proxy coefficients are highly negative and spouse coefficients are close to zero. The ASEC nonspouse coefficients reflect not only substantive underreporting, but also positive unobserved heterogeneity correlated with marriage. Among women, underreporting by proxies is small, an estimated -0.013 for annual earnings and -0.010 for hourly earnings. Misreporting by husband proxies is minimal (underreporting 1 percent or less) while nonspouse proxies underreport annual and hourly earnings by about 1.5 percent.

The substantive underreporting of men's earnings by proxies, coupled with minimal underreporting of women's earnings, has obvious implications for measurement of the gender gap, which is frequently measured using the CPS (particularly so for comparisons over long periods).[12] Proxy reports, provided most often by wives or female partners, understate men's hourly earnings by 5.5 log points. Proxy respondents (often men) understate women's hourly earnings by 1 log point, a difference of 4.5 log points. Were all earnings reported by proxies, these results would imply that the gender gap is understated by 4.5 log points. Based on sample averages of proxy use among men of 53.6 percent and 40.5 percent among women, a back of the envelope calculation implies that gender-asymmetric underreporting of

[12] Blau and Kahn (forthcoming) provide a comprehensive survey of the gender wage gap, with a focus on CPS estimates.

earnings by proxies understates the gender wage gap by about 2.5 log points [.536 x .055 – .405 x .010 = 0.0254], or about 20 percent of the adjusted wage gap. In a later section, we re-examine the gender gap in the context of quantile models.

B. Measurement Differences in the ASEC and DER

Empirical investigation of measurement error in reports of earnings in both the CPS and other surveys has a long (although sparse) history. Early investigations of earnings in the CPS include Alvey and Cobleigh (1980), Bound and Krueger (1991), and Bollinger (1998). Others have examined measurement error in earnings or hours worked in a variety of major surveys (Herriot and Spiers, 1975; Rogers and Herzog 1987; Poterba and Summers 1986; Halsey 1978; Mathiowetz and Duncan 1988; Marquis and Moore 1990; Mellow and Sider 1983; Duncan and Hill 1985; Bound et al. 1994; Bound, Brown and Mathiowetz, 2001; Roemer, 2002). While measurement error is not the focus of their paper, Bollinger and David (2001) find a relationship between nonresponse in a panel survey and measurement error in responses.

Here we briefly examine the relationship between the reported earnings in the ASEC and the earnings reported through the DER. Unlike the results in Sections 4 and 5, here we use only the linked respondents and we do not use weights. The goal of this exercise is to examine the relationship between the survey respondents for those with linked surveys (see Table 2 for descriptive statistics). Following previous work by Bollinger (1998) we use non-parametric kernel regression. Similar to the results presented in Figure 7, we first use OLS to estimate rich models of both ASEC and DER earnings on age, education, race/ethnicity, foreign born status, geography, industry and occupation. The residuals from each model are then used for the non-parametric regression of ASEC on DER.

[Figure 9 here]

The results are presented in Figure 9 using a log-earnings scale. As has been typically found in this literature, we see the "common man" hypothesis supported: individuals with low earnings tend to over-report their earnings, while individuals with high earnings tend to under-report. Since this analysis was conducted on residuals, these are not associated with demographic characteristics such as education or race. The same analysis was conducted for earnings levels (as opposed to log earnings) and for earnings absent controls (i.e., without the initial DER/ASEC earnings and log earnings regressions). Qualitative results were similar in both cases. This evidence provides some interesting qualifications on our main finding that nonresponse is concentrated in the tails of the distribution. Here we see that for respondents, measurement error is also concentrated in the tails of the distribution. Previous authors (Bollinger and David 2001; Kaptyn and Ypma 2007) have found similar overlaps in the population of "non-cooperative" survey respondents. This suggests, perhaps, that the Census imputation procedure may

reflect the response that typical nonrespondents *would make*, were they to participate, measurement error and all. It does, however, highlight that individuals in the extreme parts of the earnings distribution (both unconditional and conditional) are not responding to the survey in ways we would expect. Our prior results show that many simply do not respond, while our results here show that those who do respond, are not appropriately revealing their earnings. This adds further support to the idea that survey nonresponse and response are correlated with the level of income, even controlling for demographic factors.

C. 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. While these households are excluded from our main analysis, we do observe DER earnings for the original nonrespondent household, plus their information reported in the monthly CPS survey. However, we are unable to calculate an hourly wage measure or construct full-time full-year subsamples because we lack information on weeks and hours worked among supplement nonparticipants. Thus, we focus in this subsection on annual earnings.

[Table 8 here]

Table 8 provides limited descriptive statistics of the sample of linked whole imputes, along with the comparable full-sample means shown in Table 1. The variables chosen derive from the monthly survey data, and so are not part of the imputed ASEC record, but represent response from the individual to whom DER income was linked. It is natural to compare this group to the larger set of results shown in Tables 1 and 2. While the whole imputes are slightly younger than other nonrespondents, they have very similar race/ethnicity, gender and marital status to the group of linked nonresponders and all nonresponders. A few notable differences exist. Similar to linked groups in general, they are 83.4 percent native born. The whole imputes have lower total DER earnings, $45,587 than linked nonrespondents at $50,301. Similarly, their DER wage and salary earnings are only $43,153 compared to linked nonrespondents at $48,219. We note that in part this may attributed to higher rates of groups such as Hispanics and foreign born.

[Figure 10 here]

In Figure 10, we show how whole supplement nonresponse differs across the joint DER earnings distribution among men and women, similar to the approach seen in Figure 3 for item nonresponse. Note that the sample here differs from that seen elsewhere in the paper. The figure shows a clear-cut pattern. Supplement nonresponse (a form of unit nonresponse) is highest among those with low earnings, with a gradual and modest decline as earnings increase. In short, there is positive selection into supplement participation, with a disproportionate share of low earnings workers not participating. While the figure

has similarities to Figure 3a, the U-shape is far less evident. Given the complexity of many supplement questions, the positive selection into response may reflect cognitive difficulties and high effort costs among many lower earnings households (Kassenboehmer et al. 2015). We observe little difference in supplement nonresponse between men and women. Although supplement non-participation is only about one-half the rate of item nonresponse for earnings, positive selection into supplement participation likely leads to a modest understatement of family poverty (Hokayem et al. 2015). Recent work by Bee, Gathright, and Meyer (2015) find a similar pattern for CPS refusals. They use 1040 returns matched to respondents' addresses to examine income characteristics of households who refuse to participate in the entire survey (both the monthly and supplement). Our results for whole imputes seem to be consistent with their results, suggesting that individuals who refuse to respond to the supplement are more similar to those who refuse to participate in the survey (i.e., unit nonresponse) than those who participate but refuse the earnings question (i.e., item nonresponse).

D. Earnings Growth

One advantage of the rotation group structure of the ASEC is the overlapping nature of the sample, allowing up to 50 percent of sample individuals to be followed across adjacent ASEC years. There is a brief literature examining either measurement error or nonresponse in panel settings (Bollinger and David, 2005; Bound, Brown and Mathiowetz, 2013; Fitzgerald et al., 1998; Bound and Krueger, 1992). We briefly examine the rates of nonresponse for the two-year panels covered by our data, the relationship between earnings and nonresponse, and the impact of nonresponse on simple measures of earnings growth. A number of authors (Bollinger, 1998; Cameron and Tracy 1998; Hardy and Ziliak 2014) have pointed out that the subsample of individuals who can be followed across adjacent years in the ASEC are not fully representative because the sample frame is the household address and not the person, and thus movers are not followed. Nonetheless, the longitudinal sample is widely used and thus it is important to assess nonresponse. In our sample, we find that the linkage rate for panel individuals rises to 91.9 percent (compared to 87.4 percent for the full ASEC sample). The earnings nonresponse rate is 17.7 percent in year 1 of the panel and 18.8 percent in year 2, as compared to 21.2 percent in the full sample). As noted by Bollinger (1998), individuals who can be followed across years tend to provide more accurate reports than seen for the full sample.

[Table 9 here]

Table 9 provides means for the sample of individuals followed across years in the ASEC who are also linked to the DER. This subset is most comparable to column 4 of Table 1 (i.e., the linked sample) and the data used for much of the prior analysis. Linking the ASEC panel to the DER drops the nonresponse rate even further (similar to the full sample) to be 16.2 percent and 17.5 percent in years 1

and 2 compared to 19.6 percent for our primary ASEC sample. As noted in other papers, the panel sample is older (average age 42.9 compared to 41.5), more likely to be white (75.2 percent compared to 71.5 percent), more likely to be married with spouse present (67.7 percent compared to 58.4 percent), more likely to be native born (89.2 percent compared to 87.7 percent), and with most of the fewer foreign-born concentrated on the foreign born non-citizens (4.9 percent compared to 6 percent).

[Table 10 here]

Table 10 presents the response status in the first year cross tabulated with the response status in the second year. Overall, 73.8 percent of the sample responds in both years and 7.5 percent do not respond in either year. The joint-year response rate is of course lower than the single year response rates (83.8 percent in the first and 82.5 percent in the second year). Many individuals change their response status and such changes are approximately symmetric. We find that 10 percent of individuals respond in the first year but become nonrespondents in the second year; 8.7 percent of individuals do not respond in year 1 but then do so in year 2.

[Figure 11 here]

Figure 11 displays panel nonresponse rates plotted against the DER earnings centile for the first year in the panel. As in Panel A of Figure 3, Panel A of Figure 11 combines full-time full-year workers with part-time and part-year workers, but unlike the earlier figures, here we combine the male and female samples. The year 1 and year 2 rates are (unsurprisingly) very comparable in shape to our prior results seen in Figure 3. The third line tracks the percentage of those who failed to respond in both years. Although multi-year nonresponse is obviously lower than annual nonresponse, we again find that such nonresponse is U-shaped with respect to the level of earnings. Panel B presents the same breakdown for the full-time, full-year sample, while Panel C shows the full sample with respect to hourly wage centiles. As in comparable panels in Figure 3, we find more pronounced U-shape patterns in Panels B and C. Similar to our previous results, we conclude the nonresponse in the panel is concentrated in the two tails.

[Table 11 here]

Using IPW weights to account for individuals not linked to the DER, Table 11 presents average earnings growth between the first and second year of the panel. We focus primarily on the second column, examining the earnings growth in the DER. Overall, the average earnings growth was .017 log points. Those who responded in both years exhibited an average rate of .0142 log points. It is not surprising that the overall and complete respondent are similar given that 73.8 percent of the sample respond in both years. We note the striking pattern between those who respond only in one year: low DER earnings growth for those who respond only in year 1 and high DER earnings growth for those who respond in year two. This pattern suggests strong selection into response based on changes in earnings.

This pattern is consistent with the U-shaped pattern found in the cross-sectional analysis as well: those who have very low or very high earnings may fail to respond if that is an unusual or new situation. Earnings growth for those who fail to respond in either year is higher than those who respond in both years. This provides further evidence that nonresponse in the CPS should be treated as NMAR.

Here, unlike all other analysis, the ASEC earnings growth includes the imputations for nonresponders. We include the ASEC growth rates for comparison and evaluation of the imputation process. Comparison of growth rates between the ASEC and DER confound both measurement error and imputations in the two categories where response switches. For those who respond in both periods, measurement differences lead to ASEC having strikingly lower estimates of earnings growth. In the case of nonresponse in both periods, the ASEC imputation procedure appears to impute higher earnings growth than observed. While one can take a variety of perspectives on whether administrative earnings are the correct measure, the marked difference in relative growth suggests that the imputations may be extremely poor in capturing earnings dynamics.

7. How Troubling is Trouble in the Tails? The Consequences of Nonresponse

The linked ASEC-DER data permit us to examine directly whether relying solely on respondents' earnings produces results in a number of contexts 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 can compare estimates from respondent-only samples with those from complete samples, something not possible with publicly-available data. The CPS is a workhorse data set used throughout the social sciences largely because of its broad coverage of topics, large sample size, and long history. It is impossible for this paper to examine the implications across all relevant topics. Here we focus on three main types of estimation which should provide researchers with some guidelines for judging the importance of nonresponse in their research. In section A we examine the implications for linear models of earnings fit with least squares estimators. We find only a modest impact from using a respondent-only sample, as the symmetric nonresponse in the tails has little impact on estimation of the means. In section B, we consider the impact on coefficient estimates from quantile regressions. Here, and in particular in the lower and upper quantiles, we find estimates from respondent-only samples to be problematic compared to use of a full sample. A full sample is possible using the ASEC-DER link, but not possible with publicly available ASEC data. Our concerns regarding use of a respondent-only ASEC sample are reinforced in Section C where we examine earnings inequality. This conclusion is not surprising given that measures of inequality are sensitive to earnings in the tails.

A. Mean Earnings Estimates

Using the ASEC-DER sample and the IPW weighting to account for representativeness, we

estimate log annual earnings equations by gender, separately for the linked respondents, linked nonrespondents, and all linked workers samples, again using the dense set of covariates used previously in the analysis. Using estimates from these regressions, in Table 12 we provide the predicted earnings 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 earnings to those obtained using the coefficients from the respondent sample, which can be calculated using public ASEC data.

[Table 12 here]

Focusing first on men, use of full sample coefficients with the full sample worker attributes (X's) results in a predicted mean log earnings of 10.524. This is close to that obtained using respondent-only betas, which leads to a predicted mean log earnings of 10.534, or 0.01 (one percent) higher than obtained with the full sample. The equivalent values for women are 10.081 using full sample betas and 10.086 using respondent betas, a 0.005 difference. Here we note that we overstate mean log earnings differences between men and women by 0.6 percent using the respondent only sample. Selection is readily evident comparing predicted earnings using respondent (R) and nonrespondent (NR) betas. The R–NR predicted earnings difference is 10.534-10.489 = 0.045 for men and 10.085 – 10.059 = .026 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. In short, users of public data can avoid substantial bias by removing imputed earnings. One can rebalance the respondent sample using inverse probability weights, although in practice this rarely has a substantive effect.

The analysis comparing male and female earnings is particularly interesting because gender is the one worker attribute always matched correctly in Census imputations (Bollinger and Hirsch 2006). That is, there exists no "match bias" (i.e., wage gap attenuation) resulting from assignment of imputed earnings from a different-sex donor. Although respondent-only samples enable researchers to purge what is often severe imputation match bias, our DER analysis shows that respondent sample means can moderately differ from unobtainable full-sample means.

B. Earnings Gaps across the Distribution

We next turn to an examination of the implications of nonresponse across the *distribution* of earnings for a host of widely-studied outcomes such as earnings gaps across gender, race, and education. We examine earnings gaps at the 5th, 10th, 25th, 50th, 75th, 90th, 95th, and 99th quantiles. Figures 12-15 depict estimates of coefficients from quantile regressions of log annual earnings on the same set of

covariates used in our earlier conditional analyses. Each figure contains estimates from two samples—one using DER earnings on both matched ASEC respondents and nonrespondents (All Matched), and the other using matched respondents only (Matched Respondent). We focus here on the full-time full-year subsample in part because earnings distributions including part-time and part-year workers confound hours worked with level of earnings and thus are difficult to interpret. While wages are often used in applications, concern arises there too with differences in wage distributions between full-time/full-year workers and those who work less, as well as potential measurement error in annual hours worked. It should be noted that quantile estimates are measuring differences in quantiles of the conditional distribution, hence quantiles do not match the unconditional quantiles.

[Figure 12 here]

Figure 12 presents the estimated coefficients on the female indicator variable from pooled earnings quantile regressions from the full-time full-year subsample. The OLS coefficients (which are at the average) are presented as horizontal lines for comparison. As is typically found, overall, the male-female wage gap is larger in magnitude in the highest quantiles, and somewhat smaller in the lowest quantiles. In general, there are very few differences between the OLS estimates on the two samples (all linked and linked respondents). Respondent-only samples produce mean wage estimates highly similar to the typically unavailable full sample, thus avoiding the sometimes severe bias from including imputations. Quantile estimates at the tails, however, diverge from mean estimates. We observe gender gap estimates from the respondent-only sample that are biased in the tails. The understatement is 0.018 log points at the 5th percentile, and .021 log points at the 99th percentile, or about 10 percent of the overall gap. As noted in Figure 3, differential response rates between men and women are most pronounced in the tails of the distribution. These differential rates in the tails have little impact on averages, but estimates in the extreme tails can be problematic.

[Figure 13 here]

In Figure 13 we examine the black-white wage differential separately for men (Panel A) and women (Panel B). As in Figure 12, we see a similar pattern, where the respondent sample produces biased estimates that understate the racial gap among men. The largest impact is at the high end of the distribution, where the bias is 0.043 log points, or nearly 20 percent relative to the combined respondent-nonrespondent sample. As with the male-female differential, this is likely driven by missing high earning men. Although black men are less likely to report than white men in general, it appears that conditional on other factors such as education, nonrespondents are disproportionately white men at the highest earnings. Here, along with the consistent under-estimation of the differential in the respondent only sample, the OLS estimates display modest under-estimation as well. In Panel B, the black-white differential for

women displays a slightly different pattern. While at the higher quantiles, the respondent only sample continues to slightly understate the gap, we note that at the lower quantiles the bias is reversed, with the respondent-only subsample slightly overstating the gap.

[Figure 14 here]

In Figure 14 we present the Hispanic-white differential by quantiles for men and women in Panels A and B, respectively. As we saw for the female black-white differential, the respondent-only sample understates the differential at the highest quantiles but overstates it at the lower quantiles. For Hispanic men, the bias in the differential is most pronounced at the highest quantiles, while for women the bias is largest at the lowest quantiles. For women, the OLS Hispanic differential is somewhat overstated using the respondent sample. There is no easy remedy to correct this using public data given that imputations are not generally matched on Hispanic status.

[Figure 15 here]

Finally, Figure 15 examines the wage differential between those whose highest degree is high school (excluding GEDs) compared to high-school dropouts (Panel A) and college graduates (with that being the highest degree) compared to high school graduates (Panel B). High school returns are systematically understated using the respondent sample, particularly so in the far left tail, but with minimal understatement at the top of the distribution. The same qualitative pattern is seen for estimates of the return to college, but with a modest downward bias throughout the entire distribution (being largest at the 90th and 95th percentiles). In both schooling return cases, the respondent sample understates the return at the means (OLS).

C. Earnings Inequality

There is limited knowledge regarding how earnings nonresponse affects the measurement of inequality, and it is not readily apparent a priori how it affects inequality. One needs to identify who fails to respond, how nonresponse differs with respect to true and typically unobserved earnings (conditional on covariates), how any such nonresponse bias might differ across the earnings distribution, and how one can best treat topcoded earnings. Census uses different topcode values depending on earnings source, and these values differ between internal and public release versions of the ASEC. A key advantage of the DER data is that earnings are not topcoded, thus permitting a direct comparison of estimates of upper-tail inequality from tax records to topcoded survey responses. Some inequality studies have excluded imputed earners (Lemieux 2006; Autor, Katz, and Kearney 2008), while others have not (Burkhauser et al. 2012). This is the first such direct comparison from linked individual survey and tax data on how nonresponse and topcoding affects earnings inequality estimates. We estimate several leading measures of inequality emphasized in the recent literature—including the Gini coefficient (Figure 16), the earnings share

accruing to the top 1 percent of earners (Figure 17), and the 90/10 (Appendix Figure 1), 90/50 (Appendix Figure 2), and 50/10 (Appendix Figure 3) percentile ratios—for alternative earnings definitions.

[Figure 16 here]

In Panel A of Figure 16 we show the earnings Gini for the full sample of workers. Shown in dash-dot line is the full ASEC sample, in long-dash line is the ASEC with imputes excluded, in the solid line is the DER for all linked workers (and ASEC for non-linked), and in the short-dash line is the DER for linked respondents. Comparing the full ASEC with imputes versus ASEC respondents only, one sees that the respondent-only sample shows too low a level of inequality owing to the omission of nonrespondents disproportionately represented in the far left and right tails. Hence, omission of imputes is inappropriate for measuring unconditioned inequality. Even with IPW reweighting (in this case, reweighting based on response status), the respondent-only sample displays slightly lower dispersion than does the full ASEC with imputes. As with the ASEC, removing nonrespondents from the DER reduces the Gini measure. The larger impact in the DER reflects the fact that the imputations in the ASEC do not capture the NMAR aspect of nonresponse. IPW is unable to correct for nonrandom, nonresponse in the tails. As compared to the two DER measures, the ASEC measures show a substantially lower level of inequality and somewhat different trends. Earnings inequality in the ASEC is roughly flat over the full sample period, and everywhere below the DER. Using DER earnings, we find a higher degree of inequality (about 10 percent higher) and a modest upward trend after 2007. In the most recent years, DER display a modest upward trend, while ASEC displays a slight downward trend. Panel A establishes that NMAR nonresponse has an impact on measures of inequality. Removing those missing values will results in a downward bias in estimation, while the imputations may be failing to fully account for the nonresponse.

In Panels B and C of Figure 16, we explore whether the gap between ASEC and DER earnings inequality is due to nonresponse (Panel B) or due to differences in measurement of earnings, including topcoding (Panel C). Panel B shows three series—the ASEC inclusive of nonrespondents; the DER for linked respondents and nonrespondents (and ASEC for nonlinked); and a hybrid DER measure that uses DER earnings for linked nonrespondents and ASEC earnings for respondents (and ASEC for the nonlinked). In all cases the sample size is held constant by using the ASEC for nonlinked respondents and nonrespondents. Here we see that the hybrid measure produces a Gini level roughly halfway between the pure ASEC and DER measures. Comparing the hybrid measures to the pure ASEC measures supports the conclusion that nonrandom, nonresponse bias (NMAR) causes an understatement in the level and trend in earnings inequality based solely on ASEC.

Panel C presents the original ASEC and DER series, along with two additional series, one where the DER is used only for topcoded ASEC values with a DER link and the other where we replace the

ASEC with the DER for workers in the top but not bottom half of the ASEC earnings distribution, regardless of imputation or topcode status. The former case is of interest because the full ASEC and DER groups include a convolution of nonrespondents and topcoded workers, and thus it is less obvious what direct role the topcode in the internal ASEC plays vis-à-vis administrative tax data. The latter case is of interest because the DER does not capture earnings off-the-book, and thus the higher level of inequality observed in the DER might be an artifact of underreported earnings in the lower half of the distribution. The results in Panel C demonstrate that topcoded earnings alone in the internal ASEC are not the primary cause of the gap in inequality estimates from tax data in the DER versus ASEC survey data. The DER-only series shows substantially higher and (to a lesser extent) rising inequality as compared to ASEC earnings with DER replacing ASEC topcodes. In addition, the majority of the gap between the DER and ASEC earnings inequality arises from earnings in the upper half of the ASEC distribution, and not from off-the-books underreporting in the lower half. This conclusion is based on the minimal differences between the DER-only series and the hybrid ASEC-DER series with DER earnings replacing the ASEC in the top half of the ASEC distribution.

[Figure 17 here]

We next turn in Figure 17 to trends in the share of earnings accruing to the top 1 percent of workers in Figure 17, the most prominent inequality measure presented from tax data (Piketty and Saez 2003). Here we see clear-cut differences between the ASEC and the DER. Among all workers, there is a modest downward trend in the top 1 percent share in the ASEC, and little overall trend in the DER, but with an increase in the share in 2010. Averaged over all years, the DER measure of the top centile is 2.9 percentage points higher than the ASEC measure, or 30 percent higher than the ASEC mean share of 9.6 percent. This gap grew over time, with the DER-ASEC gap averaging 2.7 percentage points in the first half of the sample period, and 3.1 percentage points in the second half. As with the Gini, Panel B shows that nonresponse accounts for at least half of the gap between the ASEC and the DER. We show a similar story in Appendix Figures 1-3 for percentile ratios. ASEC measures of inequality tend to understate inequality because the Census hot deck (owing to nonresponse bias) imputes earnings for nonrespondents that are too high in the left tail and too low in the right tail, thus understating inequality.

8. Conclusion

This paper set out to address three questions not adequately examined in prior literature relying on a unique restricted-access dataset that links ASEC household files to administrative earnings tax records. First, how do nonresponse and patterns of nonresponse bias vary across the earnings distribution and are these patterns similar for women and men (and other groups)? Although levels of nonresponse differ based on gender, race, and ethnicity, U-shaped patterns of nonresponse across the earnings

distribution are highly similar across groups. Likewise, we see substantial differences in the level of nonresponse based on the survey month in sample and for proxy versus self-respondents, yet we see highly similar U-shaped patterns of nonresponse with respect to earnings for each of these groups. With or without conditioning on covariates, we find a U-shaped nonresponse pattern, with left-tail "strugglers" and right-tail "stars" being least likely to report earnings. Women and men have similar U-shaped nonresponse patterns across the distribution, although men have a higher level of nonresponse.

Second, is nonresponse ignorable? The short answer is no. As stated above, nonresponse is not independent of realized earnings, with or without control for covariates. Relatedly, earnings differ with respect to response status, conditional on covariates. Although nonresponse is nonignorable, throughout much of the earnings distribution there is relatively little correlation between response status and earnings. However, there is trouble in the tails.

Our third question asks if there are economic implications of nonrandom nonresponse on estimates of earnings gaps and inequality. We do find small biases at the means for some wage gaps (e.g., schooling returns and racial/ethnic wage gaps). Gender gaps are slightly understated throughout much of the distribution, but substantively understated in both the left and right tails. 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 are understated in absolute value. Race, gender, and returns to schooling gaps in the tails can be off by at least 10 percent due to nonresponse. Particularly pronounced are estimates of upper-tail inequality where nonresponse accounts for half of the 30 percent gap between survey and tax record estimates.

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 prior work, even if nonresponse were completely missing at random, severe "match bias" can arise in the estimation of earnings equation coefficients if researchers include nonrespondents whose earnings are imputed by Census. Among the "remedies" for match bias, the simplest and most widely used is to throw out imputed earnings. The respondent-only sample can be reweighted by the inverse probability of response, although in practice this typically makes little difference. This easy fix is not recommended as a response to nonrandom nonresponse as reweighting will not provide consistent estimates. This is particularly true for research focusing the upper and lower tails of the earnings distribution.

Nonresponse bias prevents researchers from observing a wide range of low earners and many high earners at the very top of the distribution. 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 in the top percentiles, a group for whom analysis already was difficult due to

reported earnings being masked (topcoded) in public use files. This difficulty is confounded by our finding that the missing at random assumption is violated, with earnings nonresponse being particularly high in the far-right tail of the earnings distribution. The Census has made some headway in addressing the topcode issue in public data with the introduction of "rank proximity swapping" whereby they now order topcoded earners from lowest to highest and randomly swap out earnings between individuals within a bounded range (and again, below the internal topcode). Unlike the previously used cell-mean series, this new approach preserves the distribution of earnings above the topcode.[13] Solving the problem of survey nonresponse, however, is much more difficult absent access to linked administrative data. Progress on this front can continue with additional efforts to link household surveys, tax records, and federal and state-level administrative data on transfers.

[13] Census has made available to the user community the rank-proximity swapped values for topcoded persons back to 1975 at https://www.census.gov/housing/extract_files/data%20extracts/income%20data%20files/. To further protect respondent confidentiality, Census rounds swapped values.

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.

Alvey, Wendy and Cynthia Cobleigh, "Exploration of Differences between Linked Social Security and CPS Earnings Data for 1972." In *Studies from Interagency Data Linkages*, report no. 11, pp. 11-18. Washington, D.C.: U.S. Department of Health, Education, and Welfare, 1980.

Arellano, Manuel and Stéphane Bonhomme "Quantile Selection Models with an Application to Understanding Changes in Wage Inequality," *Econometrica* 85 (January 2017): 1-28.

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.

Bee, C. Adam, Graton M.R. Gathright, and Bruce D. Meyer. "Bias from Unit Non-Response in the Measurement of Income in Household Surveys," Unpublished Paper, August 2015.

Blau, Francine and Lawrence Kahn. "The Gender Wage Gap: Extent, Trends, and Sources." *Journal of Economic Literature*, forthcoming.

Bollinger, Christopher R. "Measurement Error in the Current Population Survey: A Nonparametric Look," *Journal of Labor Economics* 16 (July 1998): 576-594.

Bollinger, Christopher R. and Martin H. David, "Estimation With Response Error and Nonresponse: Food-Stamp Participation in the SIPP," *Journal of Business & Economic Statistics*, 19 (April 2001): 129-141.

Bollinger, Christopher R. and Martin H. David, "I Didn't Tell and I Won't Tell: Dynamic Response Error in the SIPP," *Journal of Applied Econometrics* 20 (May/June 2005): 563-569.

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, Greg J. Duncan, and Willard L. Rodgers. "Evidence on the Validity of Cross-sectional and Longitudinal Labor Market Data," *Journal of Labor Economics* 12 (July 1994): 345-368.

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.

Cameron, Stephen, and Joseph Tracy. "Earnings Variability in the United States: An Examination Using Matched-CPS Data," Working Paper, Federal Reserve Bank of New York (1998). https://www.newyorkfed.org/medialibrary/media/research/economists/tracy/earnings_variability.pdf.

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.

Duncan, Greg J. and Daniel H. Hill, "An Investigation of the Extent and Consequences of Measurement Error in Labor-economic Survey Data," *Journal of Labor Economics* 3 (October 1985): 508-532.

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.

Halsey, H. "Validating Income Data: Lessons from the Seattle and Denver Income Maintenance Experiment," *Proceedings of the Survey of Income and Program Participation Workshop, Survey Research Issues in Income Measurement: Field Techniques, Questionnaire Design, and Income Validation*, U.S. Department of Health, Education and Welfare, Washington, D.C., 1978.

Hardy, Bradly and James P. Ziliak. "Decomposing Trends in income Volatility: The "Wild Ride" at the Top and the Bottom," *Economic Inquiry* 52 (1) (January 2014): 459-476.

Heckman, James J. "Sample Selection Bias as Specification Error," *Econometrica* 47 (January 1979): 153-161.

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.

Hirsch, Barry T. and John V. Winters. "Rotation Group Bias in Measures of Multiple Job Holding," *Economics Letters* 147 (2016): 160-163.

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

2015): 935-045.

Joe, Harry. *Dependence Modeling with Copulas*, CRC Press (Chapman and Hall), Boca Raton, FL, 1997.

Kapteyn, Arie and Jelmer Y. Ypma. "Measurement Error and Misclassification: A Comparison of Survey and Administrative Data," *Journal of Labor Economics* 25 (July 2007): 513-550.

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.

Krueger, Alan, Alexandre Mas, and Xiaotong Niu. "The Evolution of Rotation Group Bias: Will the Real Unemployment Rate Please Stand Up?" *Review of Economics and Statistics* 99 (May 2017): 258-264.

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.

Lee, Jungmin and Sokbae Lee. "Does It Matter Who Responded to the Survey? Trends in the U.S. Gender Earnings Gap Revisited," *Industrial and Labor Relations Review* 65 (January 2012): 148-160.

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.

Marquis, Kent H. and Jeffrey C. Moore. "Measurement Errors in SIPP Program Reports," in *Proceedings of the 1990 Annual Research Conference*, 721-745. Washington, DC: US Bureau of the Census, 1990.

Mathiowetz, Nancy A. and Greg J. Duncan. "Out of Work, Out of Mind: Response Error in Retrospective Reports of Unemployment," *Journal of Business and Economic Statistics* 6 (April 1988): 221-229.

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.

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

Poterba, James M. and Lawrence H. Summers. "Reporting Errors and Labor Market Dynamics," *Econometrica* 6 (November 1986): 221-229.

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.

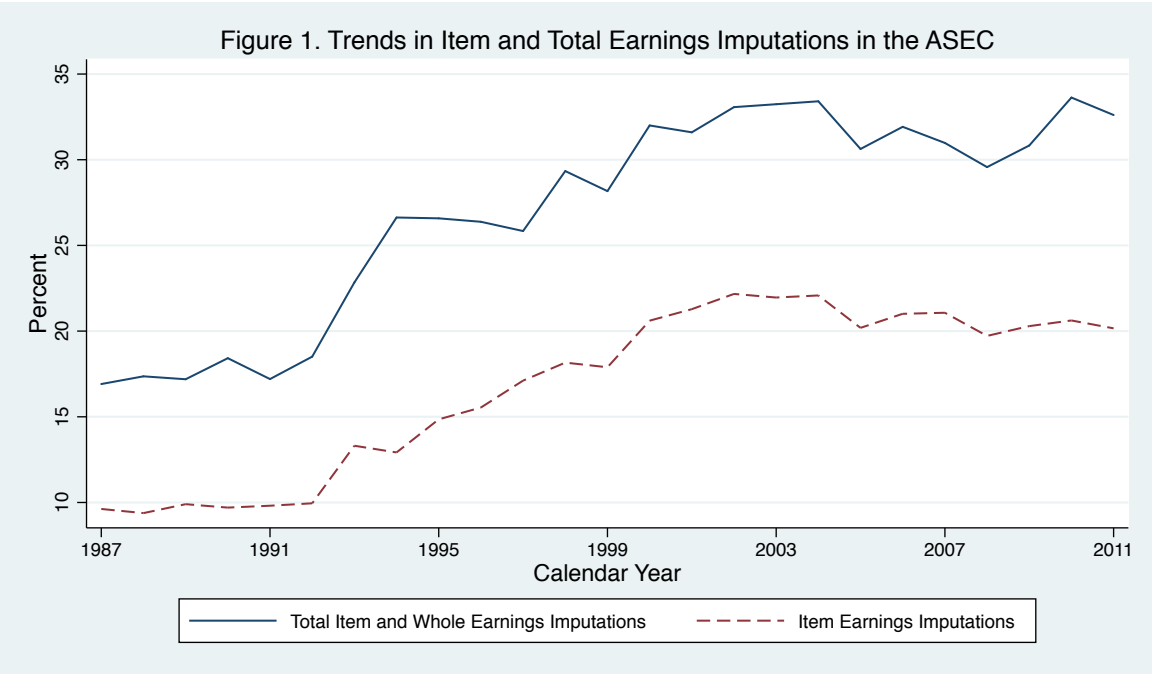
Roemer, Mark. "Using Administrative Earnings Records to Assess Wage Data Quality in the Current Population Survey and the Survey of Income and Program Participation." Longitudinal Employer-Household Dynamics Program Technical Paper No. TP-2002-22, US Census Bureau, 2002.

Rogers, Willard L. and A. Regula Herzog. "Covariances of Measurement Errors in Survey Responses," *Journal of Official Statistics* 3 (October 1987): 403-418.

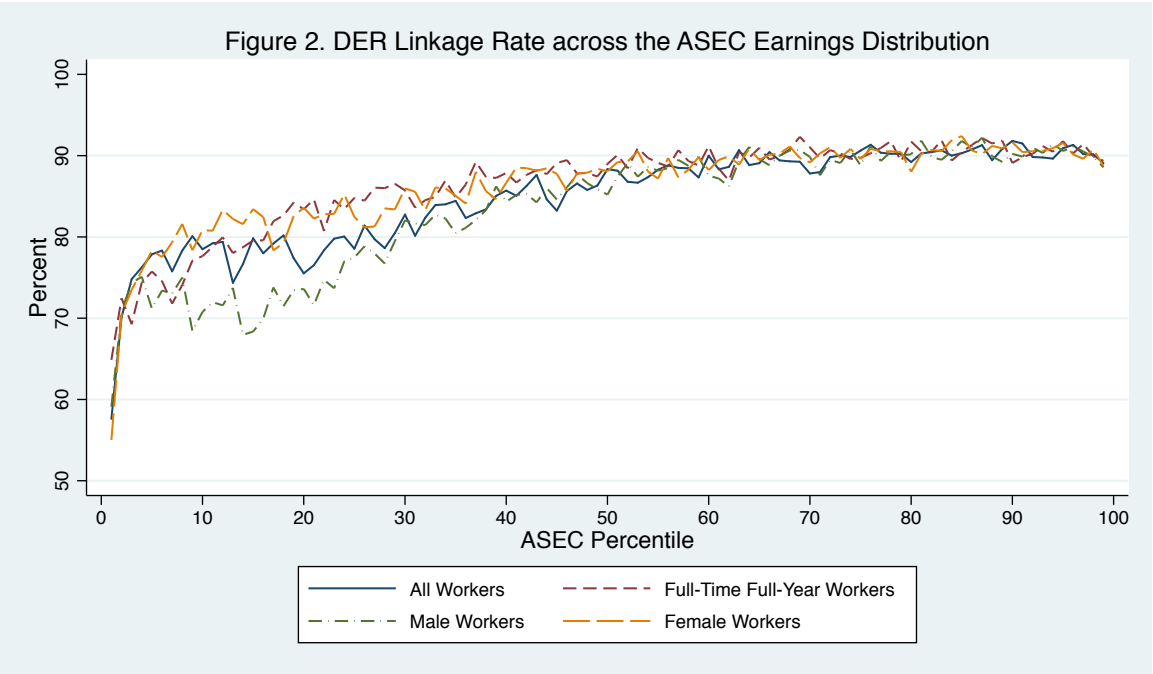
Rubin, Donald B., "Inference and Missing Data (with Discussion)," *Biometrika* 63 (1976): 581-592.

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

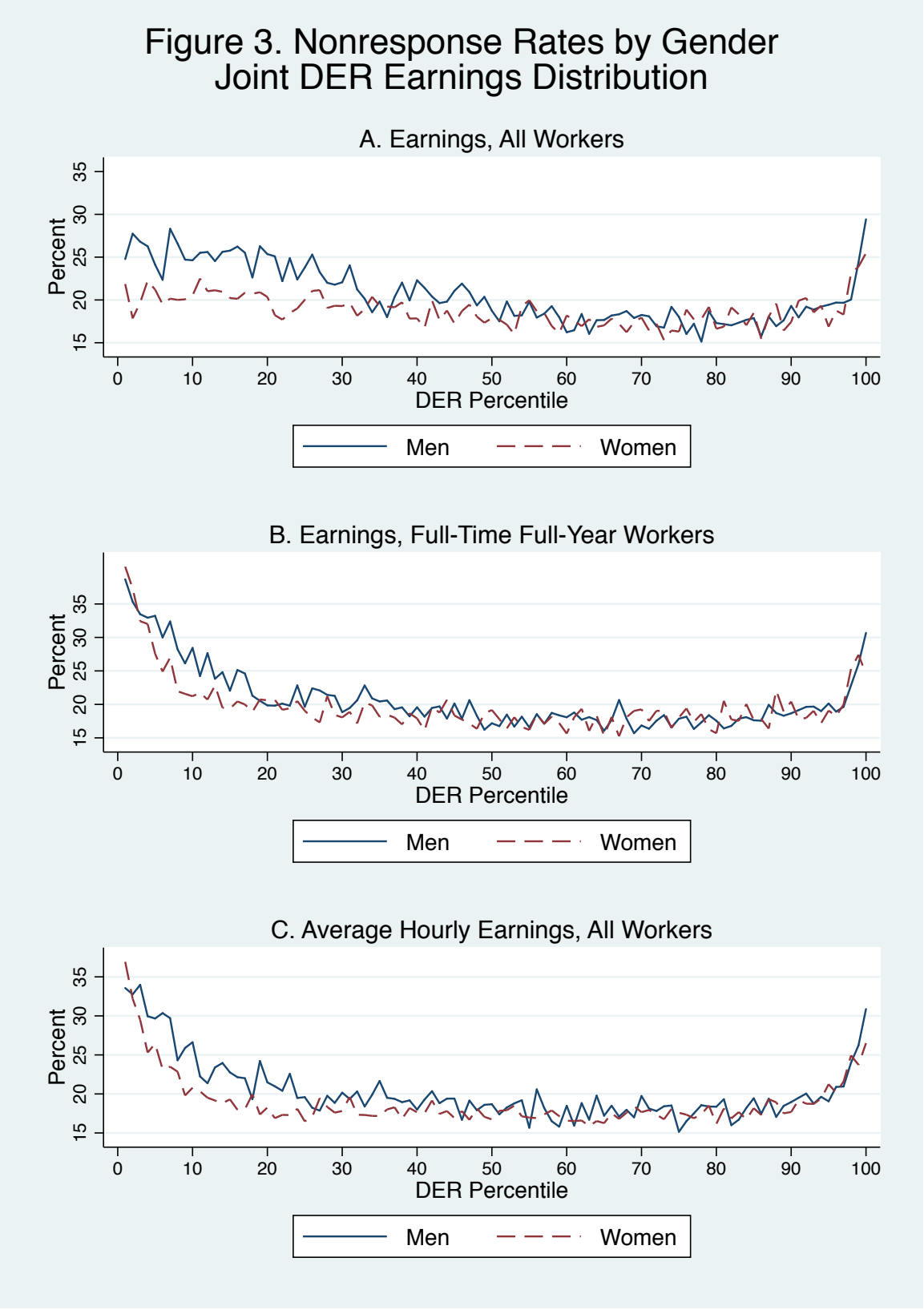
Welniak, Edward J. "Effects of the March Current Population Survey's New Processing System On Estimates of Income and Poverty," Proceedings of the American Statistical Association, 1990.



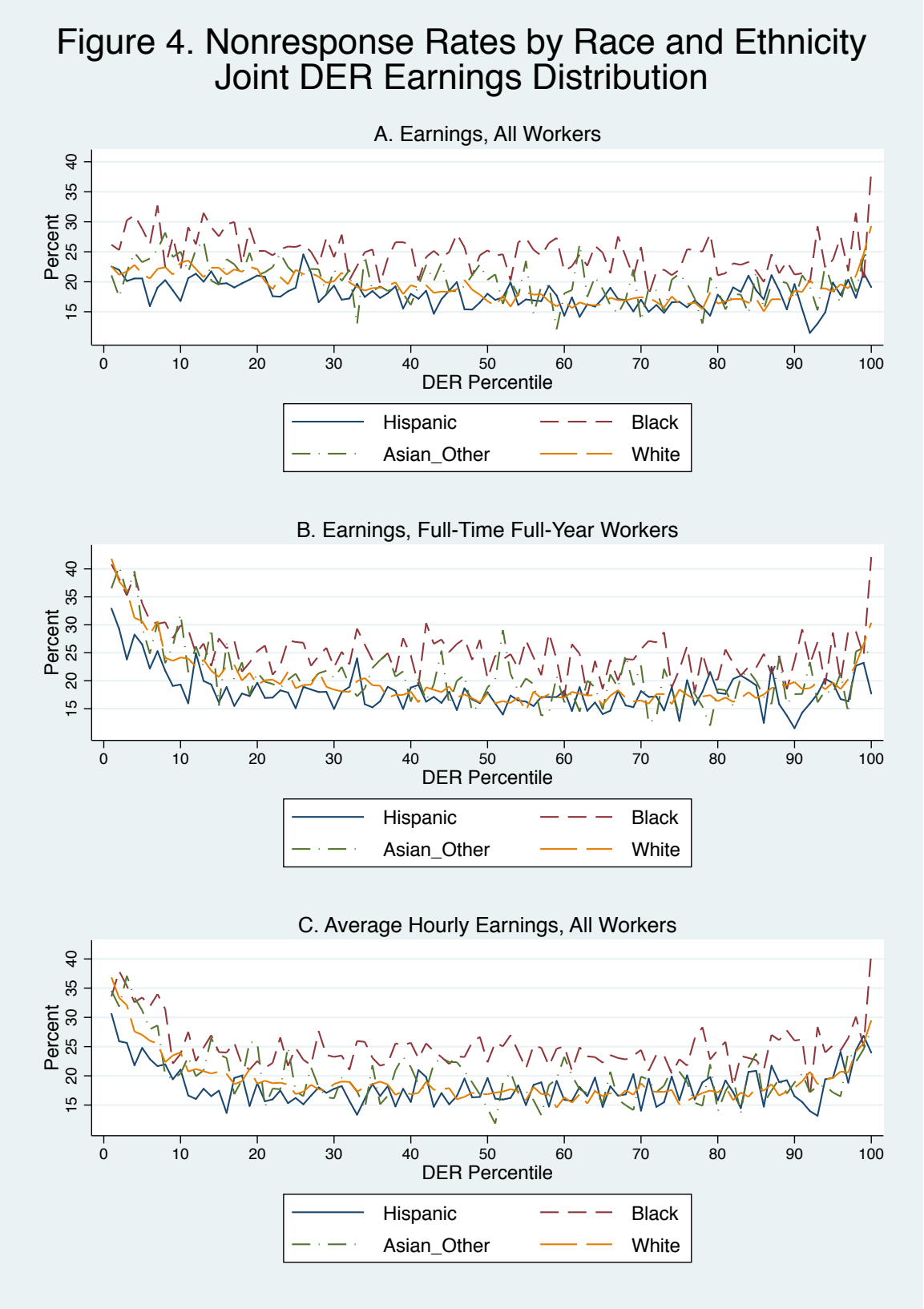
Source: Authors' calculations. U.S. Census Bureau, Current Population Survey, 1988-2012 Annual Social and Economic Supplement.



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

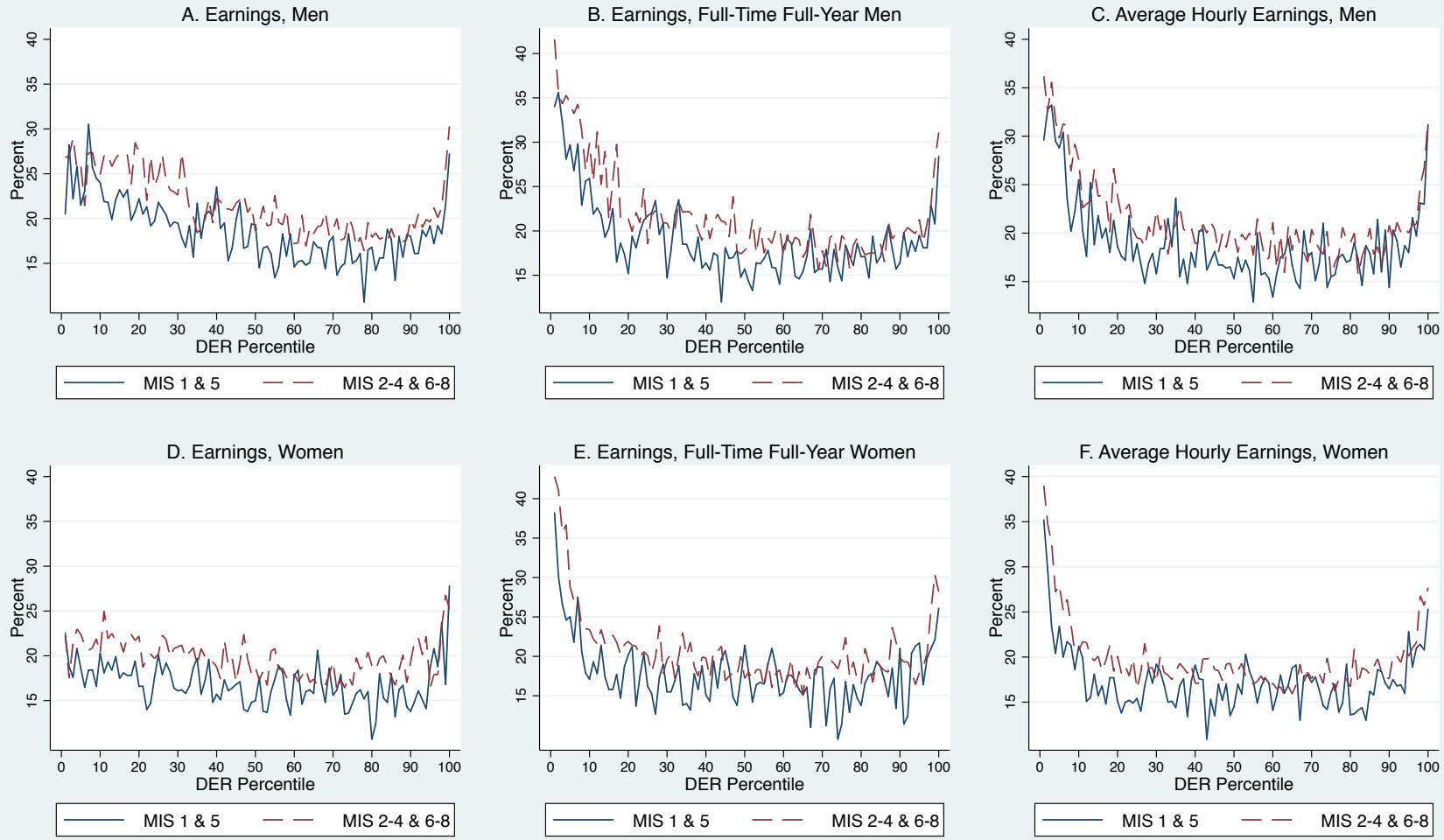


Sources: U.S. Census Bureau, Current Population Survey, 2006-2011 Annual Social and Economic Supplement. Social Security Administration, Detailed Earnings Record, 2005-2010.



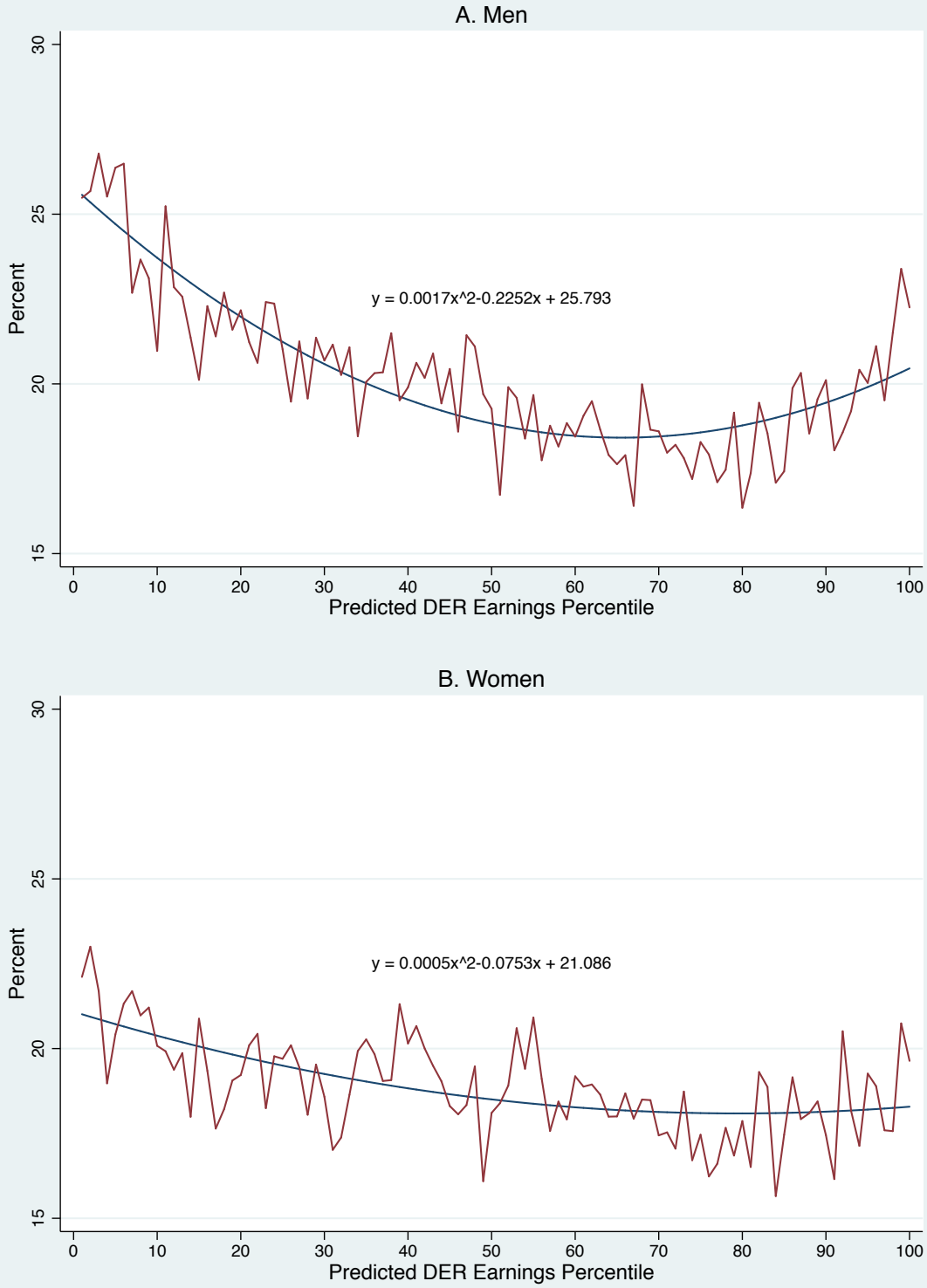
Sources: U.S. Census Bureau, Current Population Survey, 2006-2011 Annual Social and Economic Supplement. Social Security Administration, Detailed Earnings Record, 2005-2010.

Figure 5. Nonresponse Rates by Month in Sample Joint DER Earnings Distribution



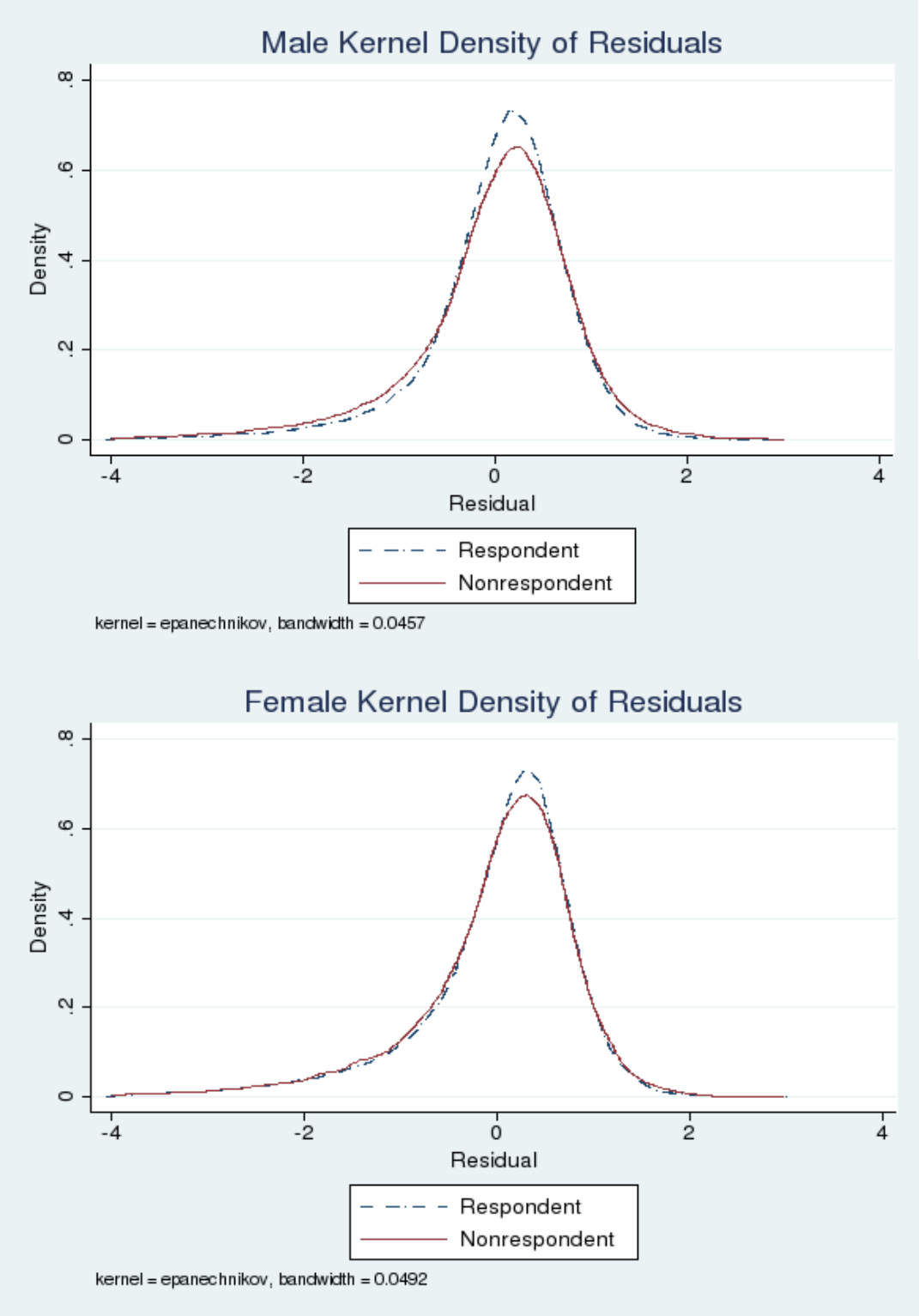
Sources: U.S. Census Bureau, Current Population Survey, 2006-2011 Annual Social and Economic Supplement. Social Security Administration, Detailed Earnings Record, 2005-2010.

Figure 6. Nonresponse Rate by Predicted DER Earnings



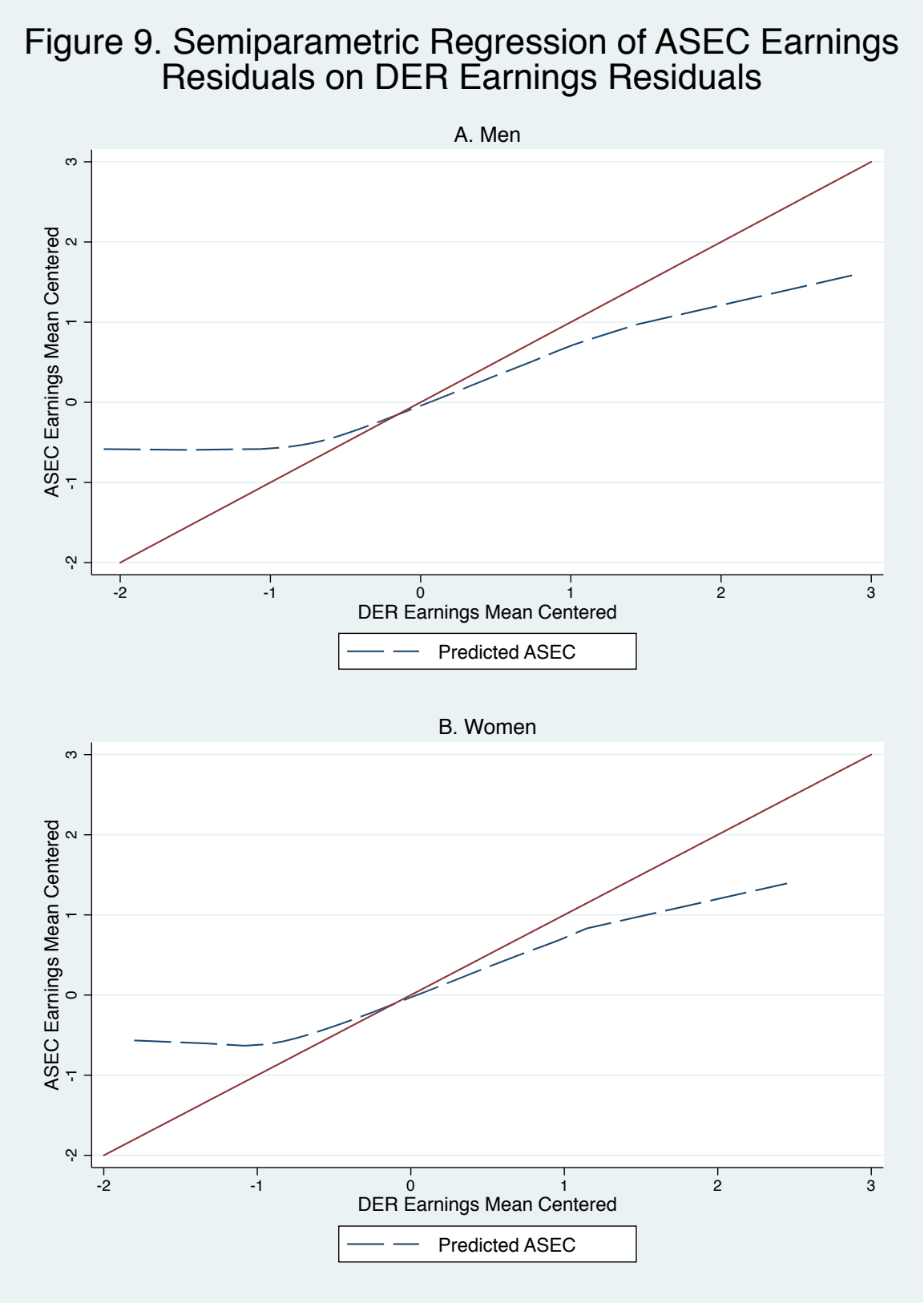
Sources: U.S. Census Bureau, Current Population Survey, 2006-2011 Annual Social and Economic Supplement. Social Security Administration, Detailed Earnings Record, 2005-2010.

Figure 7: Residuals of Log Earnings Regressions by Response Status in ASEC, by Gender

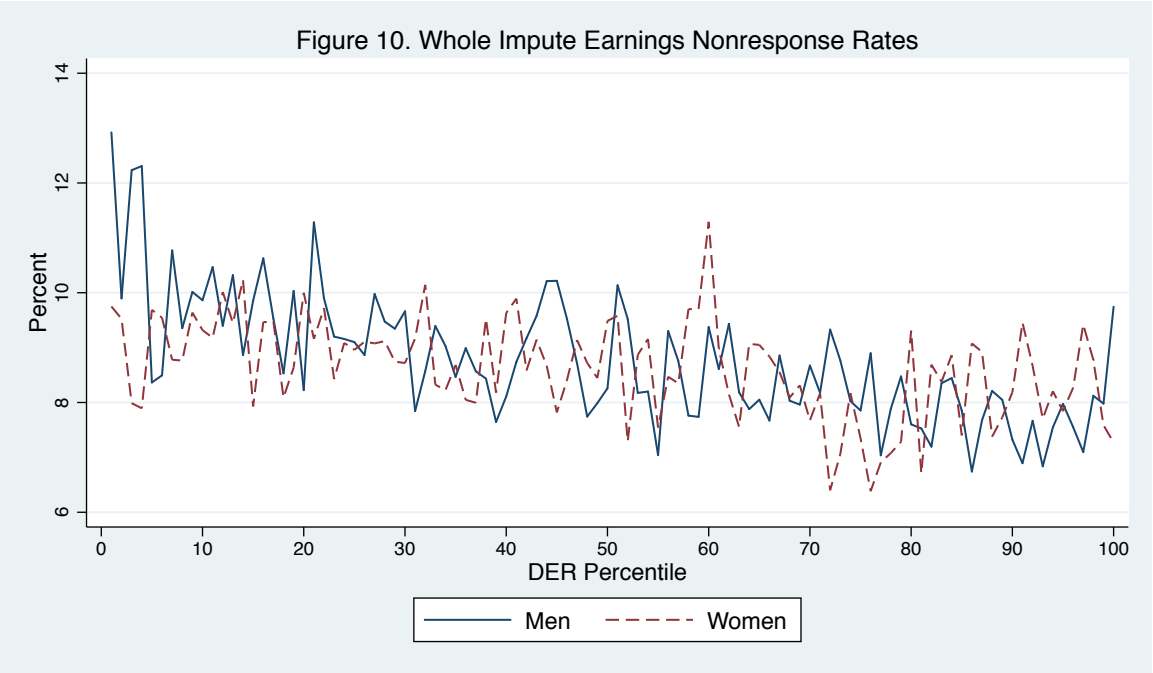


Sources: U.S. Census Bureau, Current Population Survey, 2006-2011 Annual Social and Economic Supplement. Social Security Administration, Detailed Earnings Record, 2005-2010.

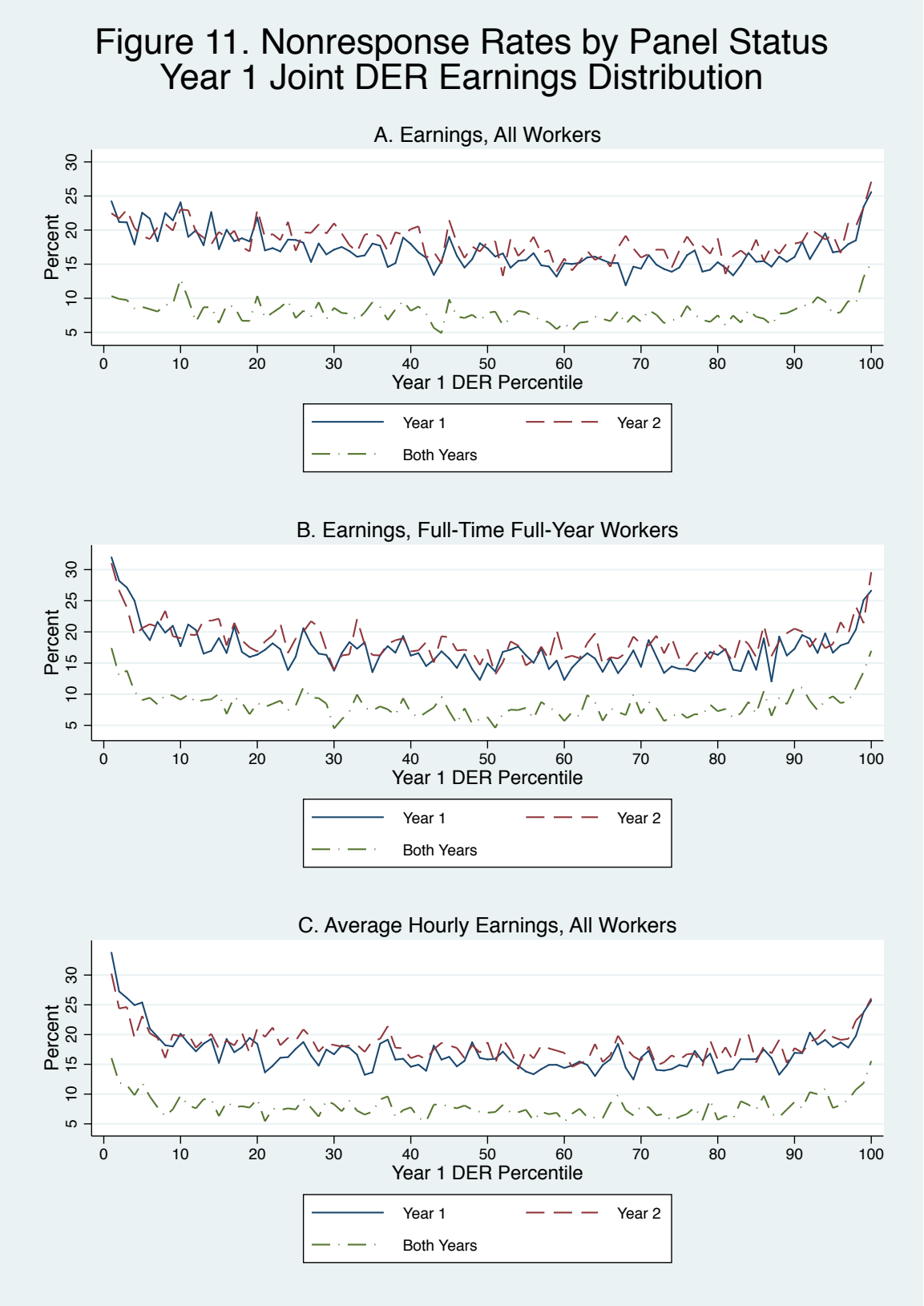
Sources: U.S. Census Bureau, Current Population Survey, 2006-2011 Annual Social and Economic Supplement. Social Security Administration, Detailed Earnings Record, 2005-2010.



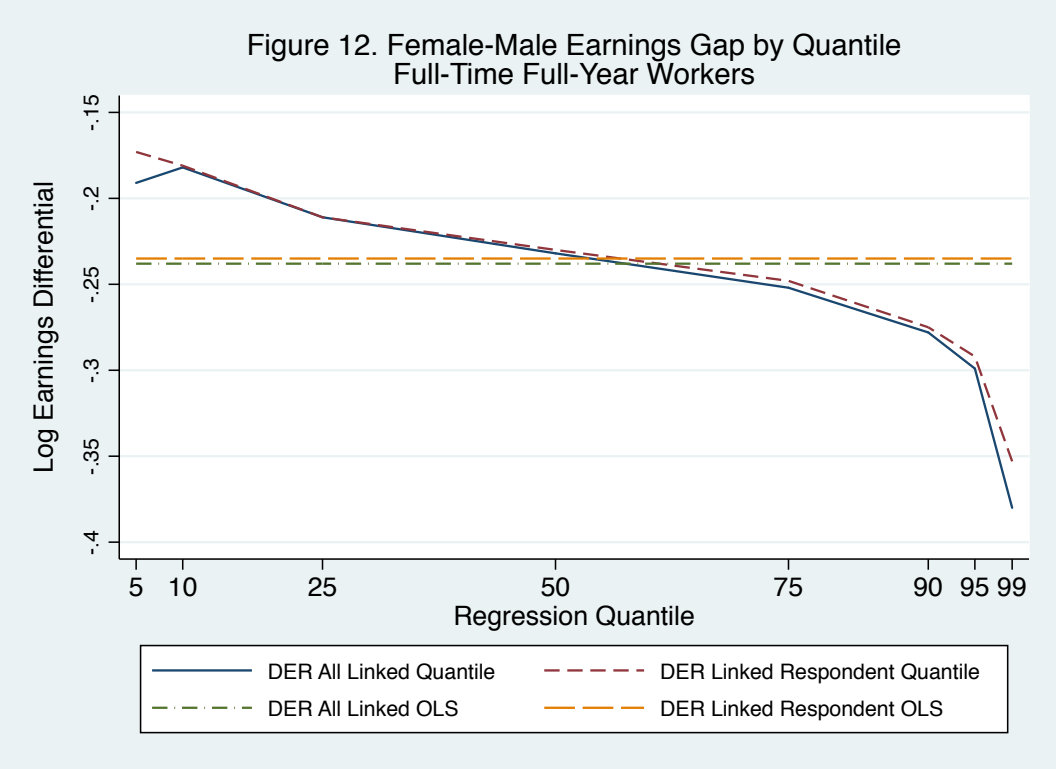
Sources: U.S. Census Bureau, Current Population Survey, 2006-2011 Annual Social and Economic Supplement. Social Security Administration, Detailed Earnings Record, 2005-2010.



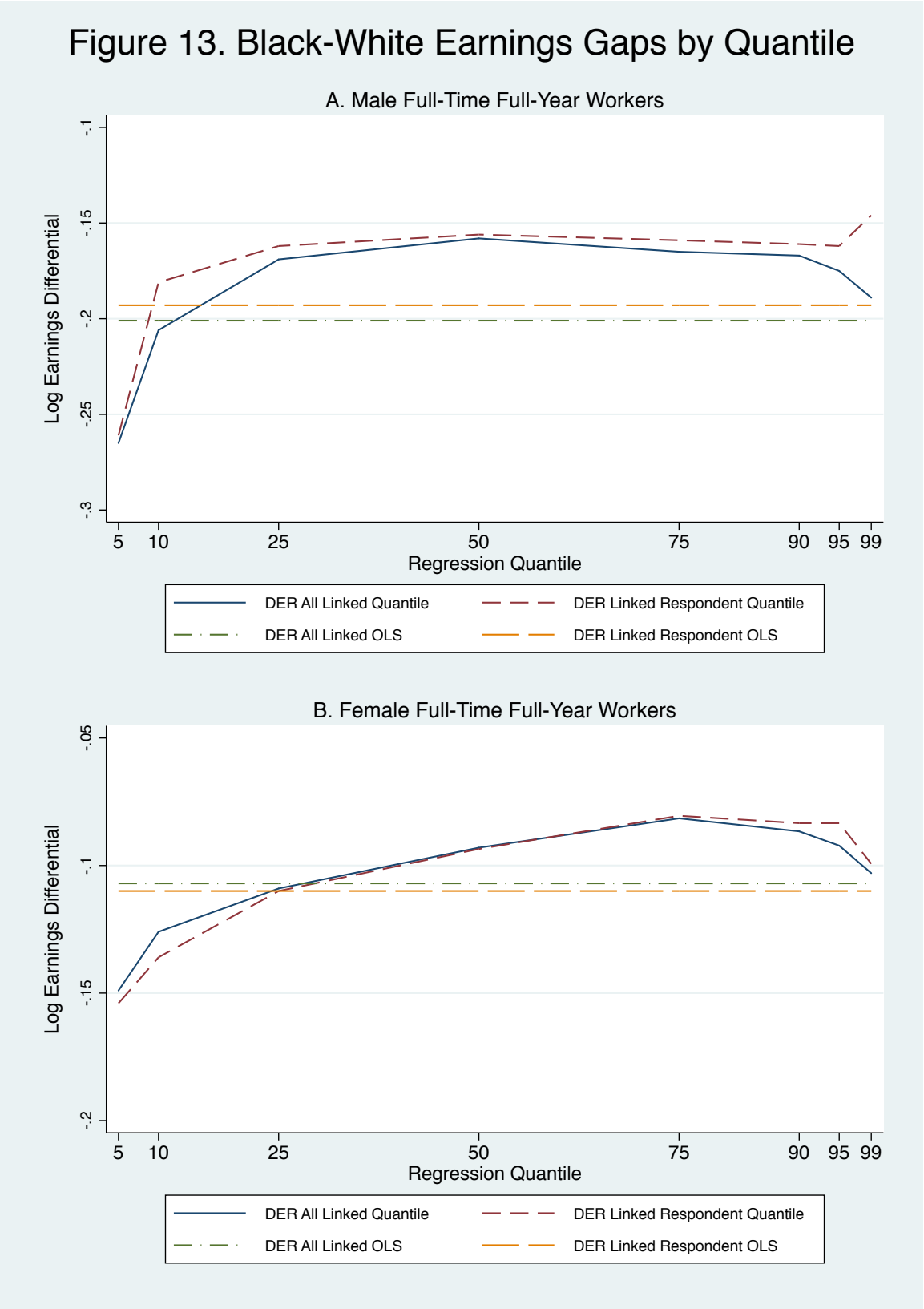
Sources: U.S. Census Bureau, Current Population Survey, 2006-2011 Annual Social and Economic Supplement. Social Security Administration, Detailed Earnings Record, 2005-2010.



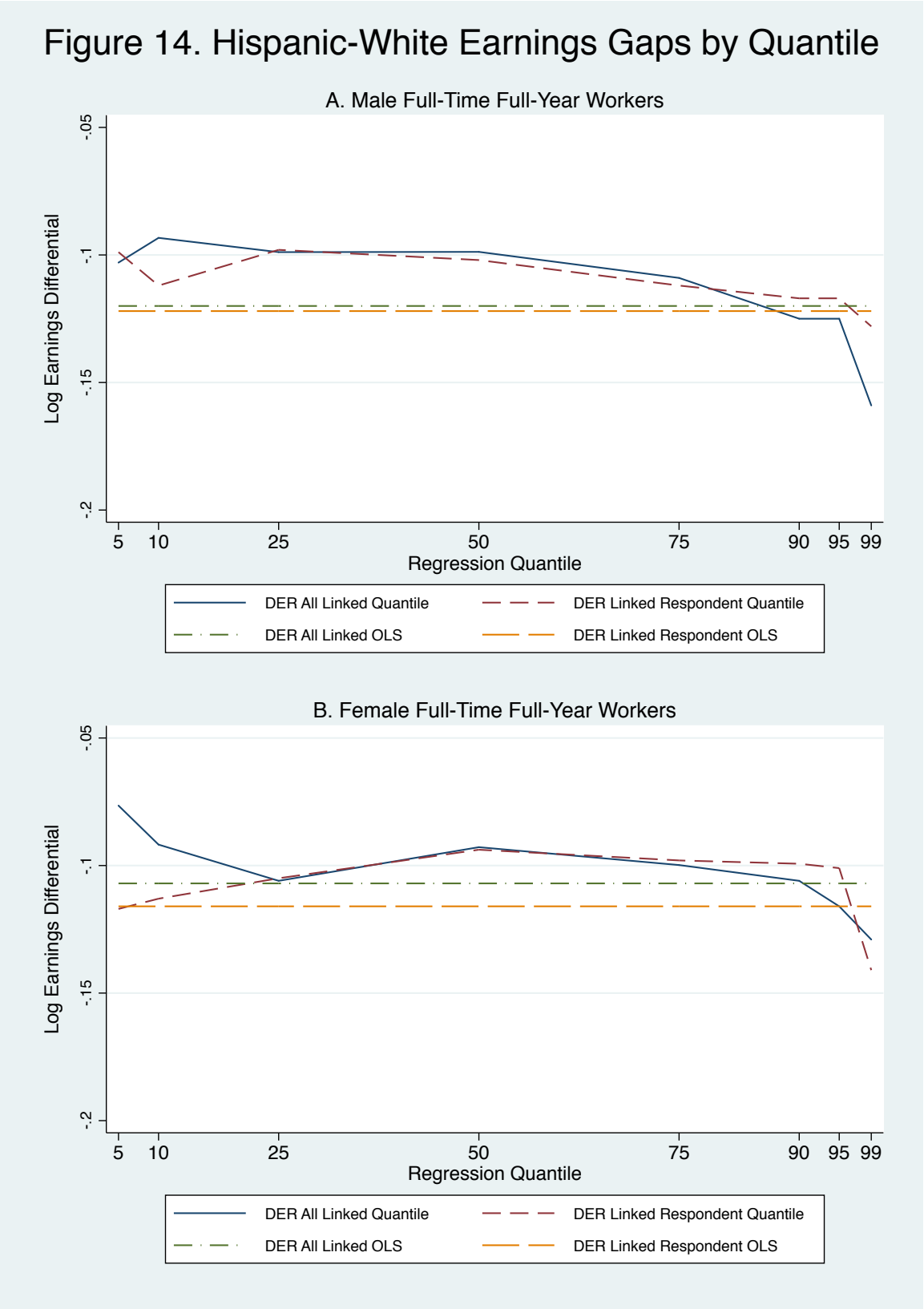
Sources: U.S. Census Bureau, Current Population Survey, 2006-2011 Annual Social and Economic Supplement. Social Security Administration, Detailed Earnings Record, 2005-2010.



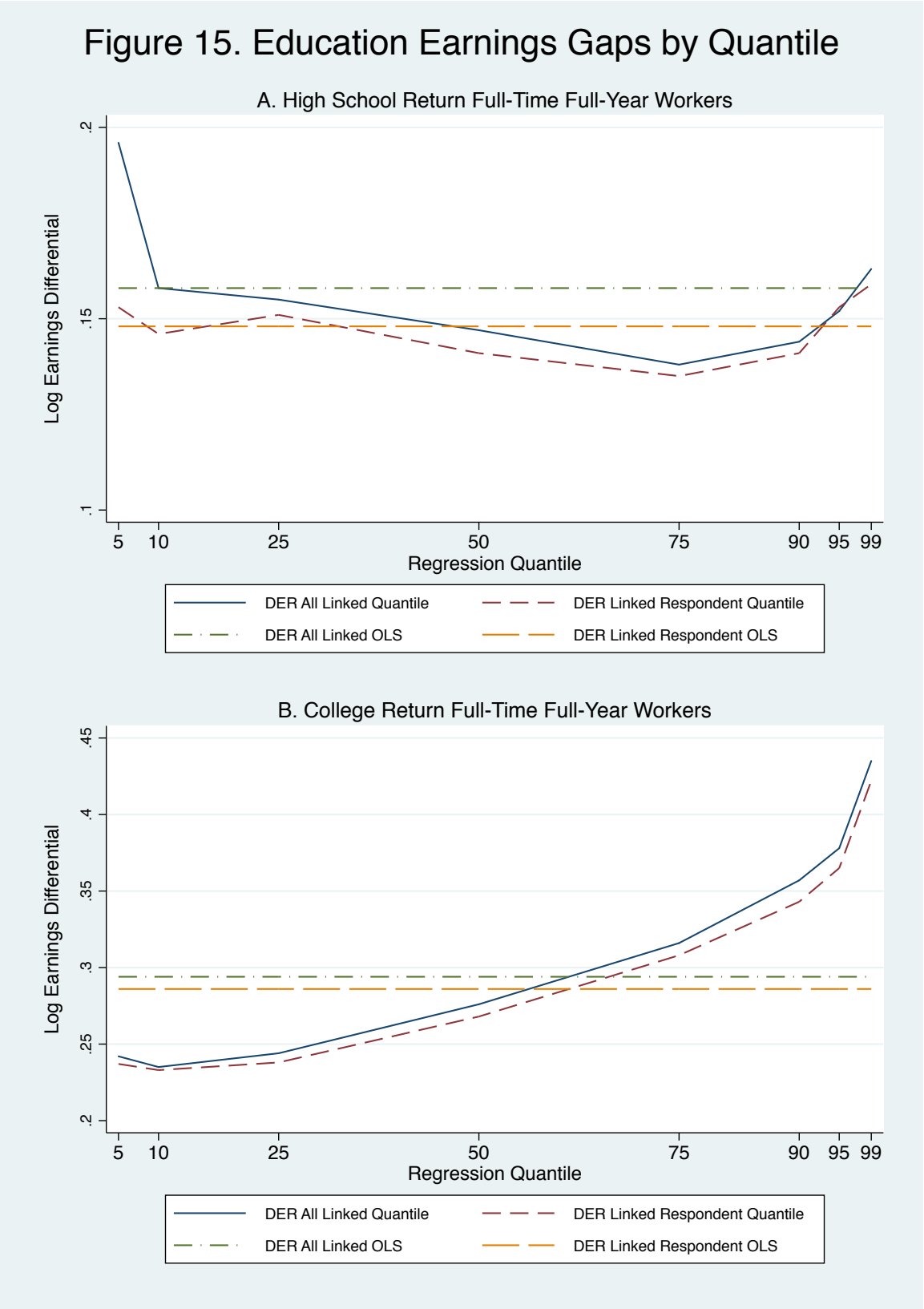
Sources: U.S. Census Bureau, Current Population Survey, 2006-2011 Annual Social and Economic Supplement. Social Security Administration, Detailed Earnings Record, 2005-2010.



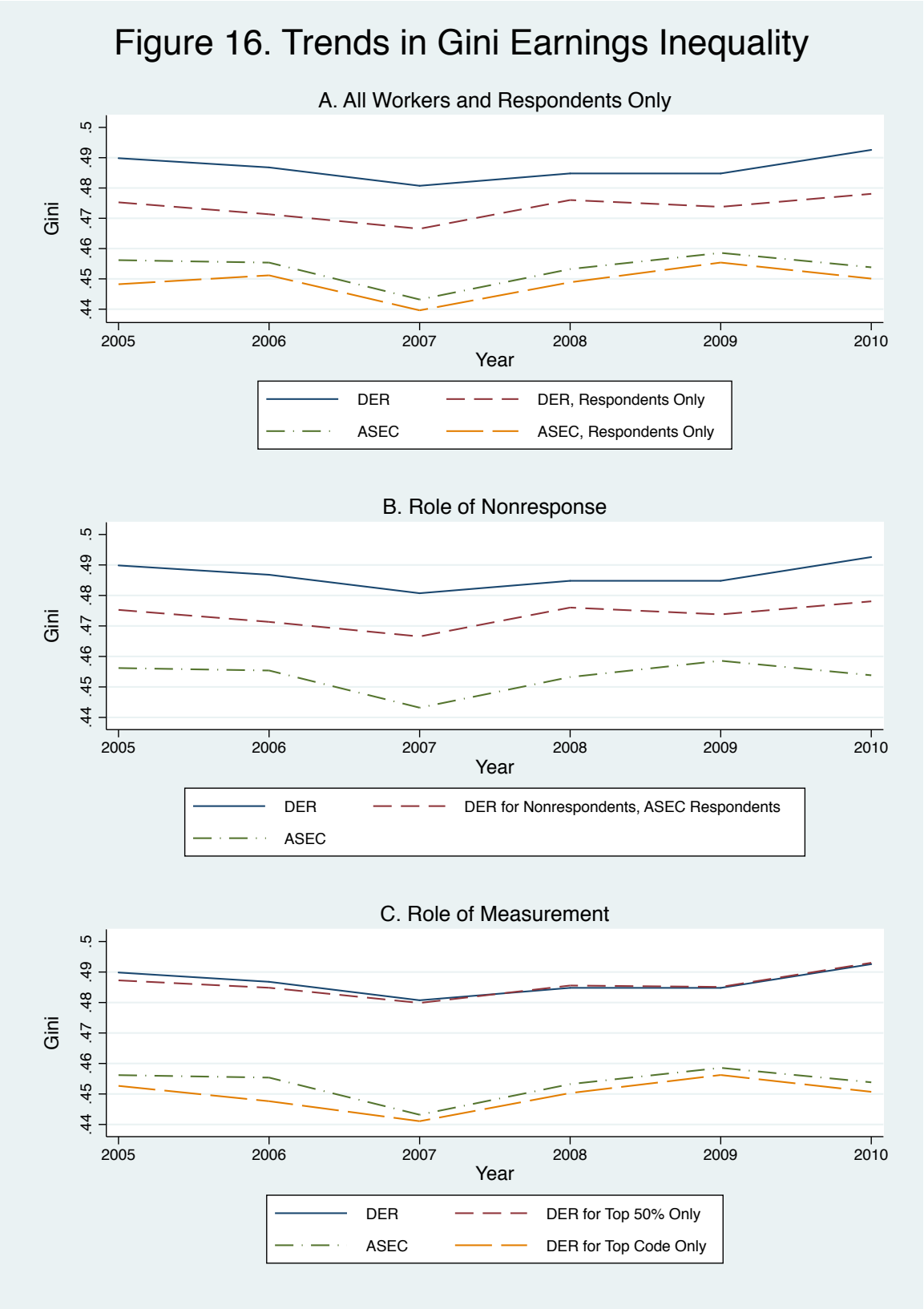
Sources: U.S. Census Bureau, Current Population Survey, 2006-2011 Annual Social and Economic Supplement. Social Security Administration, Detailed Earnings Record, 2005-2010.



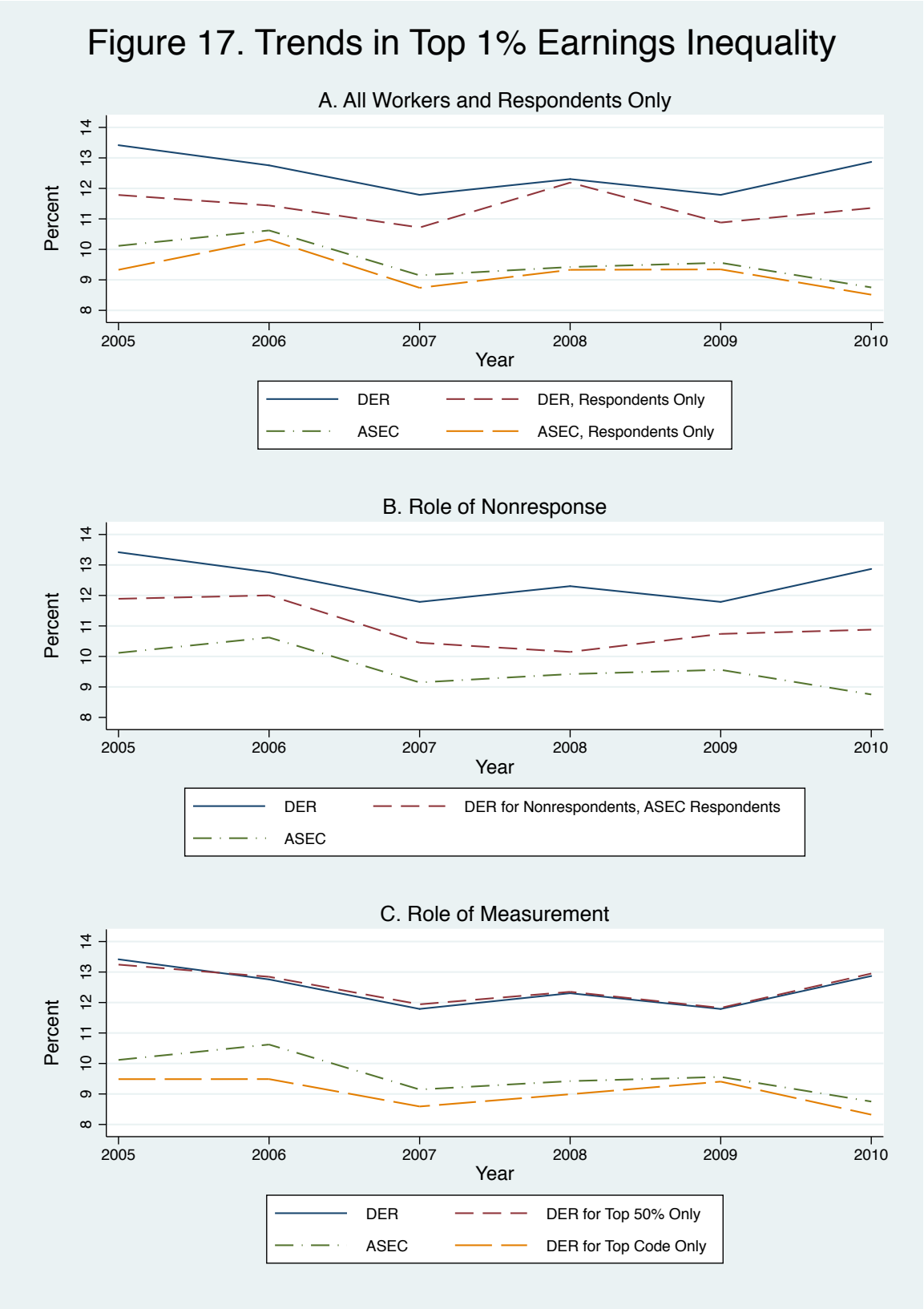
Sources: U.S. Census Bureau, Current Population Survey, 2006-2011 Annual Social and Economic Supplement. Social Security Administration, Detailed Earnings Record, 2005-2010.



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



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



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. Sample Averages by Response Status and by Linkage Status

		Response Status		Linkage Status	
	Full Sample	Respondent	Nonrespondent	Linked	Nonlinked
Age	41.1	41.0	41.4	41.5	38.7
Race/Ethnicity					
White, Non-Hispanic	68.4	69.1	65.7	71.5	46.3
Black, Non-Hispanic	10.9	10.1	13.6	10.8	11.3
Asian, Non-Hispanic	4.5	4.3	5.1	4.2	6.5
Other Race, Non-Hispanic	1.9	1.9	1.7	1.9	1.7
Hispanic	14.4	14.6	13.9	11.6	34.2
Gender					
Female	47.5	48.0	45.5	48.5	40.4
Male	52.5	52.0	54.5	51.5	59.6
Education (years)	13.7	13.7	13.6	13.9	12.5
Marital Status					
Married, Spouse Present	57.1	58.2	53.2	58.4	47.7
Married, Spouse Absent	16.8	16.8	16.8	16.6	18.4
Single, Never Married	26.1	25.0	30.0	25.0	33.9
Nativity					
Native	84.4	84.7	83.5	87.7	61.5
Foreign Born, US Citizen	6.4	6.2	7.1	6.3	7.8
Foreign Born, Not a US Citizen	9.1	9.1	9.4	6.0	30.7
Employment					
Full Time, Full Year	71.9	71.2	74.4	72.7	66.5
Full Time, Part Year	8.2	8.0	8.7	8.0	9.4
Part Time, Full Year	13.3	13.9	11.4	13.1	15.3
Part Time, Part Year	6.6	6.9	5.5	6.2	8.8
Work Hours (per week)	40.1	40.0	39.9	40.2	38.8
Nonresponse					
Nonresponse Rate (W&S or SE)	21.8	0	100	19.6	37.1
Nonresponse Rate (W&S)	21.6	0	99.0	19.4	37.0
Linkage Rate	87.4	89.9	78.6	100	0
Proxy	48.2	45.3	58.9	47.3	54.4
ASEC Total Earnings ($2010)	46,178	46,116	46,423	47,716	35,522
DER Total Earnings ($2010)	48,360	47,886	50,301	48,360	NA
ASEC W&S Earnings ($2010)	45,671	45,740	45,449	47,210	35,008
DER W&S Earnings ($2010)	46,679	46,303	48,219	46,679	NA
DER Ave. Hourly Total Earnings ($2010)	25.61	25.23	28.16	25.80	NA
Observations	479,140	372,316	106,824	405,352	73,788

Sources: U.S. Census Bureau, Current Population Survey, 2006-2011 Annual Social and Economic Supplement. Social Security Administration, Detailed Earnings Record, 2005-2010. Weighted by March Supplement Weights. Averages for ASEC include imputed nonrespondent earnings.

Table 2. Sample Averages by Link and Response Status

	Linked Respondent	Linked Nonrespondent	Nonlinked Respondent	Nonlinked Nonrespondent
Age	41.4	41.8	38.0	39.8
Race/Ethnicity				
White, Non-Hispanic	72.1	69.3	42.7	52.5
Black, Non-Hispanic	10.1	13.8	10.4	12.9
Asian, Non-Hispanic	4.1	4.4	5.9	7.4
Other Race, Non-Hispanic	1.9	1.8	1.8	1.5
Hispanic	11.8	10.7	39.2	25.7
Gender				
Female	48.9	46.6	39.7	41.4
Male	51.1	53.4	60.3	58.6
Education (years)	13.9	13.7	12.3	12.9
Marital Status				
Married, Spouse Present	59.3	54.7	47.8	47.4
Married, Spouse Absent	16.6	16.4	18.6	18.2
Single, Never Married	24.1	28.9	33.6	34.4
Nativity				
Native	87.8	87.5	57.2	68.7
Foreign Born, US Citizen	6.1	6.8	7.4	8.5
Foreign Born, Not a US Citizen	6.1	5.7	35.4	22.8
Employment				
Full Time, Full Year	72.1	75.2	63.5	71.6
Full Time, Part Year	7.9	8.4	9.2	9.7
Part Time, Full Year	13.5	11.2	17.0	12.3
Part Time, Part Year	6.5	5.2	10.3	6.4
Work Hours (per week)	40.3	40.1	38.6	39.2
Nonresponse				
Nonresponse Rate (W&S or SE)	0	100.0	0	100
Nonresponse Rate (W&S)	0	98.8	0	99.7
Linkage Rate	100	100	0	0
Proxy	44.6	58.7	51.3	59.6
ASEC Total Earnings ($2010)	47,702	47,772	32,021	41,460
DER Total Earnings ($2010)	47,886	50,301	NA	NA
ASEC W&S Earnings ($2010)	47,322	46,751	31,677	40,657
DER W&S Earnings ($2010)	46,303	48,219	NA	NA
DER Ave. Hourly Total Earnings ($2010)	25.23	28.16	NA	NA
Observations	325,903	79,449	46,413	27,375

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. Response and Linkage Rates by Hispanic Origin and Immigrant Status

	Full Sample			Native Born		Immigrants			
						Naturalized Citizen		Non-Citizen	
	All	Non-Hispanic	Hispanic	Non-Hispanic	Hispanic	Non-Hispanic	Hispanic	Non-Hispanic	Hispanic
Linkage Rate	87.4	90.4	70.7	91.1	87.9	85.8	83.0	79.1	44.8
Nonresponse Rate (W&S or SE)	21.8	22.0	21.0	21.7	20.5	26.1	20.4	23.3	21.9
Nonresponse Rate (W&S)	21.6	21.7	20.9	21.4	20.4	26.0	20.2	23.1	21.9
Respondent	78.2	78.0	79.0	78.3	79.5	73.9	79.6	76.7	78.1
Percent Hispanic	16.0	0	100.0	0	100.0	0	100.0	0	100.0
Column Relative to Other Samples									
Percent of Full Sample	100.0	84.0	16.0	76.7	7.3	4.0	2.6	3.3	6.1
Percent of Hispanic sample	-	0	100.0	-	45.8	-	16.1	-	38.1
Percent of non-Hispanic sample	-	100.0	0	91.2	-	4.8	-	4.0	-
Percent of all Immigrants	-	46.0	54.0	-	-	25.2	16.0	20.8	38.0
Percent of Naturalized Citizen sample	-	-	-	-	-	61.1	38.9	-	-
Percent of Non-Citizen Sample	-	-	-	-	-	-	-	35.4	64.6
Observations	479,140	402,692	76,448	367,401	35,003	19,303	12,306	15,988	29,139

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: ASEC Mean Nonresponse with Respect to DER Earnings for Men and Women, 2006-2011

	(1) OLS	(2) Probit Marginal Effects	(3) OLS w/X's	(4) Probit w/X's Marginal Effects
		Men		
$lnEarnings^{DER}$	-0.016*	-0.015*	-0.011*	-0.010*
	(0.001)	(0.001)	(0.001)	(0.001)
Constant	0.371*		0.398*	
	(0.012)		(0.018)	
Observations	212,708	212,708	212,708	212,708
R-squared	0.002		0.017	
		Women		
$lnEarnings^{DER}$	-0.007*	-0.007*	-0.005*	-0.005*
	(0.001)	(0.001)	(0.001)	(0.001)
Constant	0.264*		0.323*	
	(0.010)		(0.017)	
Observations	207,798	207,798	207,798	207,798
R-squared	0.000		0.015	

Robust standard errors in parentheses. * p<0.01

Estimates are weighted using inverse probability 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: ASEC Nonresponse across the DER Earnings Distribution for Men and Women, 2006-2011

	(1)	(2)	(3)	(4)
	Men		Women	
DER Earnings Deciles and Percentiles	Earnings Decile Dummies OLS	Earnings Decile Dummies and X's, OLS	Earnings Decile Dummies OLS	Earnings Decile Dummies and X's, OLS
Decile 10	0.255*	0.143*	0.202*	0.110*
	(0.00366)	(0.00921)	(0.00337)	(0.00902)
Decile 20	0.246*	0.137*	0.209*	0.114*
	(0.00382)	(0.00934)	(0.00356)	(0.00906)
Decile 30	0.209*	0.104*	0.196*	0.104*
	(0.00342)	(0.00925)	(0.00329)	(0.00895)
Decile 40	0.201*	0.0997*	0.197*	0.106*
	(0.00343)	(0.00934)	(0.00332)	(0.00898)
Decile 50	0.182*	0.0818*	0.190*	0.0997*
	(0.00333)	(0.00936)	(0.00330)	(0.00903)
Decile 60	0.178*	0.0778*	0.182*	0.0936*
	(0.00327)	(0.00939)	(0.00321)	(0.00903)
Decile 70	0.174*	0.0765*	0.178*	0.0920*
	(0.00319)	(0.00939)	(0.00319)	(0.00904)
Decile 80	0.172*	0.0753*	0.174*	0.0898*
	(0.00315)	(0.00941)	(0.00316)	(0.00909)
Decile 90	0.183*	0.0875*	0.173*	0.0891*
	(0.00320)	(0.00950)	(0.00317)	(0.00915)
Percentiles 91-95	0.194*	0.0996*	0.176*	0.0917*
	(0.00456)	(0.0101)	(0.00446)	(0.00973)
Percentile 96	0.195*	0.102*	0.200*	0.113*
	(0.0102)	(0.0137)	(0.0105)	(0.0136)
Percentile 97	0.222*	0.125*	0.187*	0.100*
	(0.0108)	(0.0141)	(0.0100)	(0.0133)
Percentile 98	0.250*	0.149*	0.184*	0.0945*
	(0.0110)	(0.0143)	(0.0100)	(0.0133)
Percentile 99	0.261*	0.159*	0.187*	0.0991*
	(0.0111)	(0.0144)	(0.0102)	(0.0134)
Percentile 100	0.314*	0.207*	0.242*	0.150*
	(0.0117)	(0.0150)	(0.0109)	(0.0141)
Observations	212,708	212,708	207,798	207,798
R-squared	0.206	0.218	0.188	0.200

Robust standard errors in parentheses. * p<0.01. Estimates are weighted using inverse probability 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 6: 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%	-3.05	-3.36	-2.97	-0.39	-3.29	-3.36	-3.27	-0.09
5%	-1.46	-1.71	-1.40	-0.31	-1.73	-1.77	-1.71	-0.06
10%	-0.88	-1.07	-0.83	-0.24	-1.07	-1.12	-1.06	-0.06
25%	-0.2	-0.36	-0.29	-0.07	-0.31	-0.35	-0.30	-0.05
50%	0.13	0.12	0.13	-0.01	0.17	0.15	0.18	-0.03
75%	0.50	0.52	0.49	0.03	0.53	0.53	0.53	0
90%	0.83	0.89	0.82	0.07	0.83	0.85	0.83	0.02
95%	1.06	1.14	1.04	0.10	1.03	1.07	1.03	0.04
99%	1.61	1.82	1.55	0.27	1.48	1.58	1.45	0.13
Mean	0.017	-0.016	0.025	-0.04	-0.01	-0.02	-0.01	-0.01
Variance	0.71	0.88	0.67	0.21	0.79	0.85	0.79	0.06
Obs	212,708	40,488	172,220		207,798	37,556	170,242	

Mean, Std Dev, and Variance use inverse probability weights. Percentiles do not use 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: Proxy Misreporting of Male and Female Annual and Hourly Earnings based on CPS-ASEC and DER Differences in Proxy Coefficients

Variable	CPS-ASEC Men	CPS-ASEC Women	DER Men	DER Women	CPS Proxy Misreport Men	CPS Proxy Misreport Women
Annual Earnings:						
Earnings equations with proxy coefficients:						
Proxy	-0.0491	0.0615	0.00975	0.0743	-0.059	-0.013
Earnings equations with spouse and nonspouse proxy coefficients						
Spouse Proxy	0.0195	0.1210	0.0778	0.1310	-0.058	-0.010
Nonspouse Proxy	-0.1880	-0.0423	-0.1280	-0.0249	-0.060	-0.017
Hourly Earnings:						
Wage equations with proxy coefficients:						
Proxy	-0.0396	0.0052	0.0152	0.0146	-0.055	-0.010
Wage equations with spouse and nonspouse proxy coefficients						
Spouse Proxy	-0.0111	0.0285	0.0413	0.0347	-0.052	-0.006
Nonspouse Proxy	-0.0985	-0.0361	-0.0389	-0.0208	-0.060	-0.015

CPS proxy misreporting estimates are calculated as the difference between the ASEC and DER proxy coefficients. See the text for explanation. The CPS-ASEC wage equations exclude imputed earners since we cannot know whether the donor's earnings were self-reported or from a proxy. The DER wage equations include the same sample.

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: Sample Averages for Linked Whole Imputes

	Mean
Age	40.8
Race	
White	64.8
Black	13.9
Asian	5.5
Other	1.9
Hispanic	13.9
Gender	
Female	48.1
Male	51.9
Education (years)	13.5
Marital Status	
Married, Spouse Present	54.5
Married, Spouse Absent	16.9
Never Married	28.6
Nativity	
Native	83.4
Foreign Born, U.S. Citizen	7.4
Foreign Born, Not a U.S. Citizen	9.3
Proxy	51.1
DER Total Earnings ($2010)	45,587
DER W&S Earnings ($2010)	43,153

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

Table 9: Averages for Panel Sample

	Mean
Age	42.9
Race	
White	75.2
Black	8.5
Asian	4.1
Other Race	2.6
Hispanic	9.6
Female	49.0
Education	14.0
Marital Status	
Married Spouse Present	67.7
Married Spouse Absent	14.7
Never Married	17.6
Nativity	
Native Born	89.2
Foreign Born Citizen	5.9
Foreign Born non-Citizen	4.9
Response Rates	
Nonresponse Year 1	16.2
Nonresponse Year 2	17.5
Sample Size	97651

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

Table 10: Joint Distribution of Response Status Across Years

		Response Status in Year 2		
		Nonresponse	Response	Total
Response Status Year 1	Nonresponse	7.5%	8.7%	16.2%
	Response	10.0%	73.8%	83.8%
	Total	17.5%	82.5%	100.0%

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

Table 11: Log Earnings Growth By Response Status

	Log Earnings Growth	
	ASEC	DER
Full Sample	0.005	0.017
Nonrespondent in Both years	0.033	0.029
Respondent in Both years	0.001	0.014
Respondent only in year 1	-0.010	-0.010
Respondent only in year 2	0.026	0.062

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

Table 12: lnW^{DER} Earnings Equation Predicted Log Earnings with Full Sample, ASEC Respondents, and ASEC Nonrespondents, 2006-2011

VARIABLES	(1) Betas from lnW^{DER} All Workers	(2) Betas from lnW^{DER} Respondents	(3) Betas from lnW^{DER} Nonespondents
Men			
Prediction with full sample X's	10.524	10.534	10.489
Prediction with respondent sample X's	10.536	10.545	10.501
Observations	212,708	172,220	40,488
R-squared of earnings equation	0.346	0.349	0.339
Women			
Prediction with full sample X's	10.081	10.085	10.059
Prediction with respondent sample X's	10.086	10.091	10.064
Observations	207,798	170,242	37,556
R-squared of earnings equation	0.281	0.284	0.271

Regression estimates use inverse probability 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. 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.

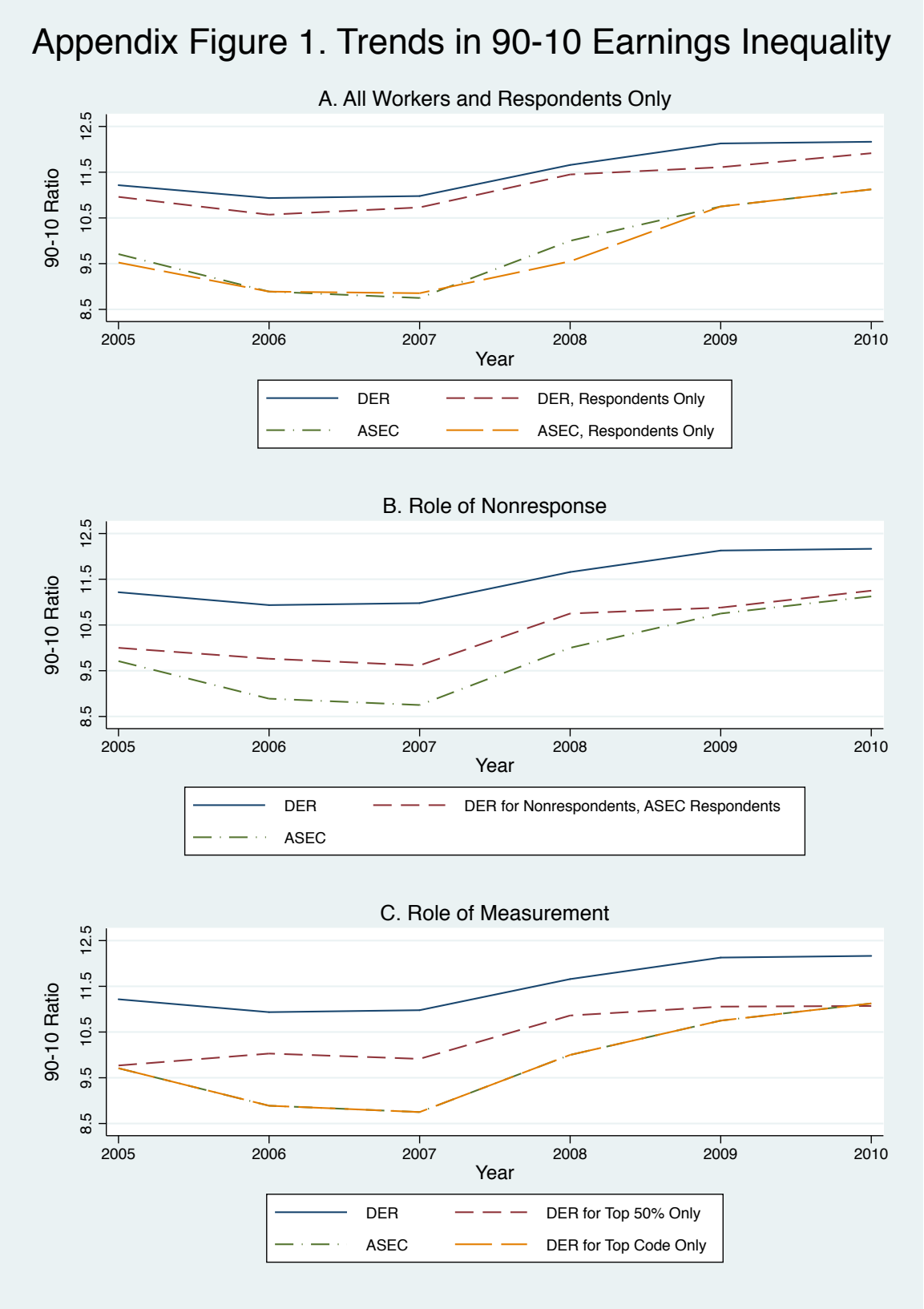
Appendix Table 1: Probit Estimates of the Probability of an ASEC-DER Link

VARIABLES	
Potential Experience	-0.0198***
	(0.00386)
Potential Experience ^{2}	0.00121***
	(0.000331)
Potential Experience^3	-2.28e-05**
	(1.08e-05)
Potential Experience^4	1.23e-07
	(1.17e-07)
Black, Non-Hispanic	-0.164***
	(0.00886)
Asian, Non-Hispanic	-0.221***
	(0.0141)
Other Race, Non-Hispanic	-0.171***
	(0.0142)
Hispanic	-0.0841***
	(0.0101)
Married-Spouse Present	0.174***
	(0.00692)
Married-Spouse Absent	0.0634***
	(0.00872)
Foreign-born Citizen	0.0709***
	(0.0139)
Foreign-born Non-Citizen	-0.499***
	(0.0123)
Foreign-born Hispanic	-0.467***
	(0.0162)
Female	0.00562
	(0.00583)
Less than 9 years schooling	0.0458**
	(0.0196)
10 years of schooling	0.184***
	(0.0208)
11 years of schooling	0.266***
	(0.0195)
12 years of schooling, no diploma	0.192***
	(0.0226)
GED	0.316***
	(0.0199)
High School Graduate	0.243***
	(0.0130)
Some College	0.349***
	(0.0140)
Associates Degree	0.388***
	(0.0153)
Bachelors Degree	0.329***
	(0.0146)
Masters Degree	0.349***
	(0.0171)
Professional Degree	0.316***
	(0.0256)
Doctorate	0.305***
	(0.0268)

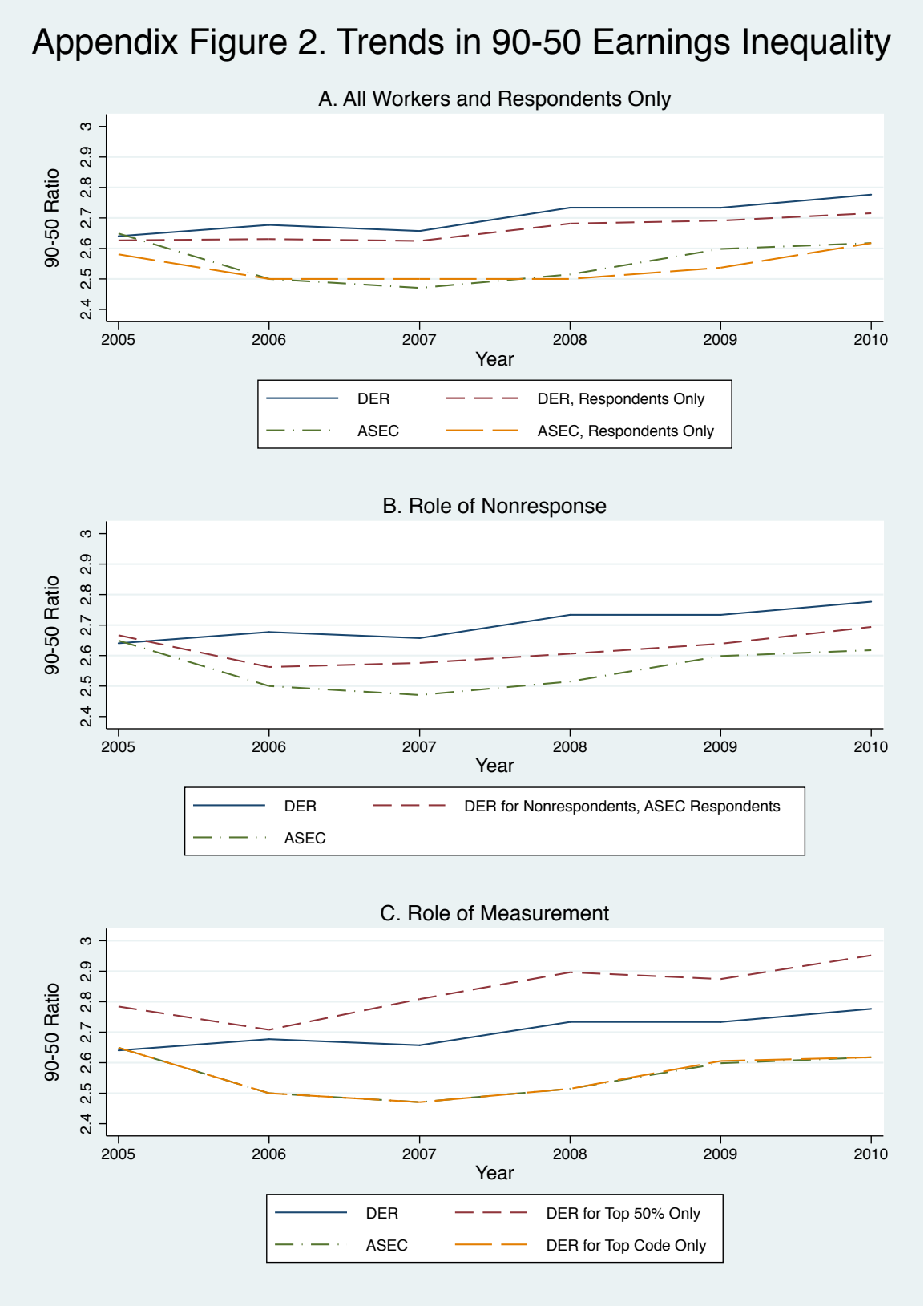
Metro size < 100k	0.184***
	(0.0122)
Metro size 100k-249k	0.217***
	(0.0119)
Metro size 250k-499k	0.183***
	(0.0122)
Metro size 500k-999k	0.160***
	(0.0124)
Metro size 1m-2.49m	0.135***
	(0.0113)
Metro size 2.5m-4.99m	0.150***
	(0.0115)
Metro size >= 5m	0.0385***
	(0.0122)
Occupation-Professional	0.0234**
	(0.00993)
Occupation-Services	-0.120***
	(0.0104)
Occupation-Sales	-0.160***
	(0.0118)
Occupation-Office Support	0.0190*
	(0.0106)
Occupation-Farm	-0.0174
	(0.0325)
Occupation-Construction	-0.162***
	(0.0155)
Occupation-Installer	0.0328**
	(0.0159)
Occupation-Production	0.00454
	(0.0135)
Occupation-Transportation	-0.0341**
	(0.0136)
Occupation-Federal (non-USPS)	-0.0461***
	(0.0172)
Occupation-Federal (USPS)	0.0676*
	(0.0379)
Occupation-State	0.0881***
	(0.0138)
Occupation-Local	0.0694***
	(0.0110)
Industry-Agriculture	-0.511***
	(0.0276)
Industry-Mining	0.100***
	(0.0337)
Industry-Construction	-0.329***
	(0.0145)
Industry-Wholesale & Retail Trade	-0.0292***
	(0.0113)
Industry-Transportation & Utilities	-0.0841***
	(0.0145)
Industry-Information	-0.0299
	(0.0191)
Industry-Finance	-0.0992***
	(0.0129)
Industry-Professional	-0.171***
	(0.0110)
Industry-Education	-0.00265

	(0.0102)
Industry-Arts	-0.177***
	(0.0119)
Industry-Other	-0.384***
	(0.0132)
Census Division-Mid Atlantic	-0.0422***
	(0.0127)
Census Division-East North Central	-0.0340***
	(0.0117)
Census Division-West North Central	0.0809***
	(0.0121)
Census Division-South Atlantic	-0.148***
	(0.0106)
Census Division-East South Central	-0.112***
	(0.0149)
Census Division-West South Central	-0.0591***
	(0.0123)
Census Division-Mountain	-0.189***
	(0.0115)
Census Division-Pacific	-0.240***
	(0.0108)
Year-2006	0.00210
	(0.00875)
Year-2007	-0.0430***
	(0.00866)
Year-2008	-0.0635***
	(0.00863)
Year-2009	-0.106***
	(0.00860)
Year-2010	-0.0613***
	(0.00881)
Constant	1.164***
	(0.0258)
Observations	479,141

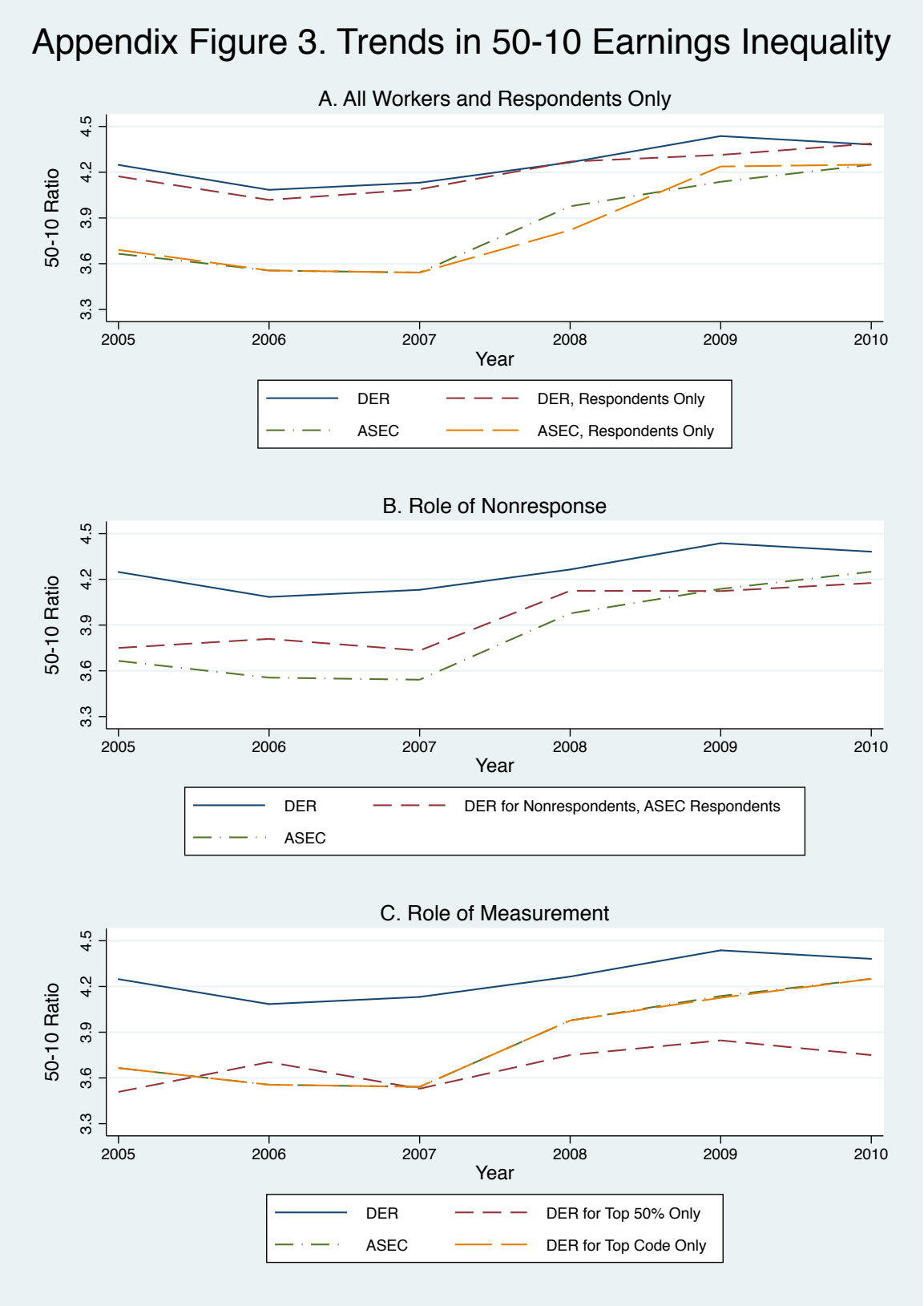
Sources: U.S. Census Bureau, Current Population Survey, 2006-2011 Annual Social and Economic Supplement. Social Security Administration, Detailed Earnings Record, 2005-2010.



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



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



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