Trouble in the Tails? Earnings Nonresponse and Response Bias across the Distribution Using Matched Household and Administrative Data

Christopher Bollinger, Barry Hirsch, Charles Hokayem, and James Ziliak

Preliminary Draft, October 2013

Abstract: This paper uses internal Current Population Survey (CPS-ASEC) individual records matched to administrative SSA data on earnings for calendar years 2004-2008 to examine the patterns and consequences of earnings non-response. Three questions not adequately addressed in prior literature are examined. First, we examine how non-response varies across the earnings distribution, a difficult question to answer absent information on non-respondents' earnings. Second, we ask whether response bias is ignorable; that is, whether respondents and non-respondents have equivalent earnings, conditional on covariates, throughout the earnings distribution. And third, we examine whether proxy responses, which account for half of all CPS earnings reports, are reliable. Our preliminary findings include the following. Non-response across the earnings distribution, conditional on covariates, is U-shaped, with left-tail "strugglers" and right-tail "stars" being least likely to report earnings. Women have particularly high non-response in the left tail; men have high non-response in the far right tail. Throughout much of the distribution (roughly the 20th through 95th percentiles) there is little correlation between response and earnings, implying that non-response is largely ignorable over this range, but with possible trouble in the tails. Proxy response is correlated with earnings, conditioning on covariates, but this largely reflects unmeasured worker heterogeneity and not misreporting of earnings.

Key words: CPS ASEC, hot deck imputation, non-response bias, earnings, measurement error

JEL Codes: J31 (Wage Level and Structure)

Submitted for presentation at the Society for Labor Economists (SOLE) Nineteenth Annual Meetings, Arlington, VA, May 2-3, 2014.

*Contact: Charles Hokayem, U.S.Census Bureau, SEHSD, HQ-7H168, 4600 Silver Hill Rd, Washington, DC 20233-8500. E-mail: charles.hokayem@census.gov Phone: 301-763-5330. Christopher Bollinger and James Ziliak, University of Kentucky. Barry Hirsch, Georgia State University. The views expressed in this research, including those related to statistical, methodological, technical, or operational issues, are solely those of the authors and do not necessarily reflect the official positions or policies of the Census Bureau. The authors accept responsibility for all errors. This paper is released to inform interested parties of ongoing research and to encourage discussion of work in progress. This paper reports the results of research and analysis undertaken by Census Bureau staff. It has undergone more limited review than official publications.

1. Introduction

Household surveys typically have high rates of earnings (and income) non-response. For example, the Current Population Survey Annual Social and Economic Supplement (CPS-ASEC) and the American Community Survey (ACS) have non-response rates on annual earnings of close to 20%. The CPS monthly outgoing rotation group earnings files (CPS-ORG) have earnings non-response rates of about 30%. Among households that do report earnings in these surveys, half the earnings reports are from a "proxy" respondent (often a spouse). Individuals for whom earnings are not reported have their earnings "allocated" using hot deck imputation procedures that assign to them the earnings of a "similar" donor who has reported earnings. Because the matching of donor earnings to non-respondents is imperfect, inclusion of imputed earners in wage analyses can introduce severe "match bias" in wage gap estimates. Simple remedies exist, but these make the assumption that non-response is ignorable.

Despite the high rates of non-response to earnings questions in household surveys, we have limited knowledge regarding three fundamental questions. First, how does non-response vary across the earnings distribution? This is difficult to know absent information on non-respondents' earnings. Second, is response bias ignorable? That is, do respondents and non-respondents have equivalent earnings, conditional on covariates? And third, how does response bias differ across demographic groups (by gender, race, marital status, citizens versus non-citizens, etc.) and with respect to different types of jobs (by occupation and industry). Progress in answering these questions will make it possible to provide guidance to researchers who use public use data files to analyze earnings determination.

In this paper, we address each of the questions above using CPS-ASEC household files matched to administrative earnings records for the March 2005-2009 (corresponding to calendar years 2004-2008, largely prior to substantial impacts from the Great Recession). Although we cannot provide fully conclusive answers to these difficult questions, we believe we make substantial progress in addressing them in ways that are highly informative. In what follows, we first provide background on each of these issues, followed by discussion of the methods used to address them, description of the matched CPS-DER data, and presentation of evidence to answer each question.

2. Background: Earnings Non-response, Imputation Match Bias, and Response Bias

Official government statistics and most researchers analyzing earnings (and income) differences include both respondents and imputed earners in their analyses, assuming (usually implicitly) that no systematic bias results. Such an assumption is often unwarranted. For analyses of earnings or wage differentials common in the social sciences, inclusion of workers with imputed earnings frequently causes a large systematic, first-order bias in estimates of wage gaps with respect to wage determinants that are not imputation match criteria or are matched imperfectly in the hot deck procedure.

This so-called "match bias" (Hirsch and Schumacher 2004; Bollinger and Hirsch 2006) occurs even when non-response is missing at random. For example, wage differentials with respect to such attributes as union status, industry, location of residence, foreign-born, etc. are severely attenuated in typical analyses. Estimates using full samples roughly equal the weighted average of largely unbiased estimates from the respondent sample and of severely biased estimates close to zero among the non-respondent (imputed) sample. For example, the full sample union-nonunion log wage gap estimate for men of 0.142 shown by Bollinger and Hirsch is roughly the weighted average of the 0.191 estimate

among earnings respondents and the 0.024 estimate among those with imputed earnings (Bollinger and Hirsch 2006, Table 2). The intuition is simple. Among those for whom earnings are imputed, most union workers are assigned the earnings of a nonunion worker; among nonunion workers, some are assigned the earnings of union workers. Absent a strong correlation between union status and attributes included in the hot deck match, the union-nonunion wage differential in the imputed sample will be close to zero. A more complex bias pattern occurs with respect to the earnings determinants that are included in the hot deck match but grouped into broad categories (e.g., schooling, age, occupation, etc., with gender providing the only exact match), leading to imperfect matches between earnings donors and non-respondents.

Although match bias can be substantial and of first order importance, it is easy to (largely) eliminate. Among the remedies are: simply exclude imputed earners from the analysis; exclude the imputations and reweight the sample by the inverse probability of response; retain the full sample but adjust estimates using a complex correction formula; or retain the full sample but conduct one's own earnings imputation procedure using all earnings covariates in one's model. In practice, each of these approaches eliminates first-order match bias and produces highly similar results (Bollinger and Hirsch 2006). Each of these methods, however, assumes earnings are missing at random (MAR) so that response bias is ignorable. That is, conditional on measured covariates, those who do and do not respond to the earnings questions would exhibit no difference in earnings, if these could be observed.[1]

Unfortunately, the validity of the MAR assumption is difficult to test. One approach is estimation of a selection model, as in Bollinger and Hirsch (2013), but such an approach relies on existence of an exclusion variable(s) that predicts non-response but is not correlated with earnings (conditional on controls), as well as reliance on distributional assumptions that cannot be directly verified. Using CPS survey methods or time period as exclusion variables (these measures affected response rates but not earnings), Bollinger and Hirsch concluded that there exists response bias, with positive selection into non-response (i.e., those with higher earnings, conditional on covariates). The bias appeared to be larger for men than for women. Importantly, they concluded that bias was largely a fixed effect that showed up in wage equation intercepts, but had little discernible effect on estimated slope coefficients.[2]

A more direct approach for determining whether or not non-response is ignorable, the approach taken in this study, is to conduct a validation survey in which one compares CPS household earnings data with administrative data on earnings provided for both CPS earnings respondents and non-respondents. There are several well-known validation studies comparing earnings information reported in household surveys with earnings recorded in administrative data. But typically these studies include only workers reporting earnings in the household survey and do not examine the issue of response bias (e.g., Mellow and Sider 1983; Bound and Krueger 1991; for a survey see Bound, Brown, and Mathiowetz 2001).

We are not the first study to examine response bias using a validation study, but prior studies examining CPS non-response are quite old, use small samples, and examine restricted populations (e.g., married white males). Most similar to our analysis is a paper by Greenlees et al. (1982), who examine the March 1973 CPS and compare wage and salary earnings the previous year with 1972 matched income tax

[1] Note that inclusion of non-respondents with imputed earnings in the estimation sample, while potentially introducing severe match bias, does not correct for response bias since the donor earnings assigned to non-respondents are drawn from the sample of respondents. The earnings of non-respondents are not directly observed.
[2] This latter conclusion was based on a comparison of wage equation coefficients from their full-sample selection models and those from OLS models in which imputed earners were excluded.

records. They restrict their analysis to full-time, full-year male heads of households in the private nonagricultural sector whose spouse did not work. They conclude that nonresponse is not ignorable, being negatively related to earnings (negative selection into response). This conclusion is based on a regression of response on administrative earnings, which yields a negative sign, conditioning on a selected number of wage determinants. The authors estimate a wage equation using administrative earnings as the dependent variable for the sample of CPS respondents. Based on these estimates they impute earnings for the CPS nonrespondents. Their imputations understate administrative wage and salary earnings of the nonrespondents by .08 log points. The sample included 561 non-respondents and earnings were censored at $50,000.[3]

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

Although informative and suggestive, it is not known whether results from these early studies examining response bias can be generalized outside their time period and narrow demographic samples. The sequential hot deck procedure used in the March survey at that time was primitive, failing to use education as a match variable (Lillard et al. 1986).[4] In short, there exists little validation evidence regarding CPS response bias with recent data. Given the increasing rates of non-response over time, it is important to know whether non-response is ignorable and, if not, the size and patterns of bias.[5]

3. The CPS ASEC Imputation Procedure for Earnings

The Census Bureau has used a hot deck procedure for imputing missing income since 1962. The current system has been in place with few changes since 1989 (Welniak 1990). The CPS ASEC uses a sequential hot deck procedure to address item non-response for missing earnings data. The sequential hot deck procedure assigns individuals with missing earnings values that come from individuals ("donors") with similar characteristics. The hot deck procedure for the CPS ASEC earnings variables relies on a sequential match procedure. First, individuals with missing data are divided into one of 12 allocation groups defined by the pattern of non-response. 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.

[3] Herriot and Spiers (1975) earlier reported similar results with these data, the ratio of CPS respondent to IRS earnings being 0.98 and of CPS imputed to IRS earnings being 0.91.
[4] Welniak (1990) provides an overview of changes over time in Census hot deck methods for the March CPS.
[5] Korinek, Mistiaen, and Ravallion (2007) examine potential bias from unit rather than item nonresponse. There is a separate literature that considers various methods to deal with missing data. These (very useful) methods, which often require strong distributional assumptions, shed little light on whether or not CPS non-response is ignorable.

For example, suppose the set of match variables consists of gender, race, education, age, and region where education is defined by less than high school, high school, some college, and college or more. If no match is found using this set of match variables, then the race variable could be dropped and education could be redefined by collapsing education categories to high school or less, some college, and college or more. If no match exists, then region could be dropped to obtain a match. This process of dropping and redefining match variables continues until the only match variable remaining is gender. This sequential match procedure always ensures a match.

The sequential hot deck used in the CPS-ASEC is a variant of a cell hot deck procedure, but rather different from the cell hot deck used in the CPS monthly outgoing rotation group earnings files (CPS-ORG).[6] Unlike the CPS-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 non-respondents 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 CPS-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 CPS-ASEC similar to that used in the CPS-ORG would not be feasible. On the one hand, reasonably detailed matching would require reaching back many years in time to find donors. On the other hand, 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 CPS ASEC also uses a hot deck procedure for unit non-response. In this context unit non-response, or a whole imputation, refers to an individual who does not respond to the ASEC supplement and requires the entire supplement to be imputed. Instead of 12 allocation groups, the whole imputation procedure uses 8 allocation groups. The set of match variables is smaller than the set used for item non-response, consisting of variables from the monthly CPS. To be considered a donor for whole imputations an ASEC respondent has to meet a minimum requirement. The requirement is at least 1 person in the household has answered one of the following questions: worked at a job or business in the last year; received federal or state unemployment compensation in the last year; received supplemental unemployment benefit in the last year; received union unemployment or strike benefit in the last year; or lived in the same house one year ago. Similar to the sequential hot deck procedure for item non-response, 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.

Unit non-response (i.e., whole imputes) is about 10%. Looking ahead, households who did not participate in the CPS-ASEC supplement have their earnings included in the matched administrative

[6] For a description of cell hot deck categories used in the CPS-ORG files over time, see Bollinger and Hirsch (2006).

earnings data described below. However, we do not observe their household characteristics since it is the donor household that is included in the CPS. For this reason, whole imputes are excluded in subsequent analysis. What we do know is that workers in households who did not participate in ASEC have lower administrative earnings than does the average earner. Because we cannot condition on individual characteristics, such evidence does not allow us to make inferences regarding non-ignorable earnings response bias.

4. Data Description: The CPS-DER Earnings Match Files

The data used in our analysis are Current Population Survey (CPS) person records matched to Social Security Administration earnings records. The CPS files used are the Census internal CPS Annual Social and Economic Supplement (CPS ASEC) data for survey years 2005-2009 (reporting earnings for calendar years 2004-2008). In addition to the data included in CPS public use files, the internal file has top-coded values for income sources that are substantially higher than the public use top codes.[7]

The Census internal CPS ASEC is matched to the Social Security Administration's (SSA) Detailed Earnings Record (DER) file. The DER file is an extract of SSA's Master Earning File (MEF) and includes data on total earnings, including wages and salaries and income from self-employment subject to Federal Insurance Contributions Act (FICA) and/or Self-Employment Contributions Act (SECA) taxation. Only positive self-employment earnings are reported in DER (Nicholas and Wiseman 2009) because individuals do not make SECA contributions if they have self-employment losses. The DER file contains all earnings reported on a worker's W-2 forms. These earnings are not capped at the FICA contribution amounts and include earnings not covered by Old Age Survivor's Disability Insurance (OASDI) but subject to the Medicare tax. Unlike ASEC earnings records, the DER earnings are not capped. This is important given that there are substantial concerns regarding non-response and response bias in the right tail of the distribution, but knowledge on these issues is quite limited. That said, in this initial draft, we cap DER annual earnings at $2 million to avoid influence from extreme earnings on wage equation coefficients.

The DER file also contains deferred wages such as contributions to 401(k), 403(b), 408(k), 457(b), 501(c), and HSA plans. The DER file does not provide a fully comprehensive measure of gross compensation. Abowd and Stinson (forthcoming) 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 IRS and the SSA. In our analysis, we can compare how discrepancies between CPS earnings reports (which are likely to include undocumented earnings) and the administrative data change in samples with and without demographic or industry-occupation groups of workers most likely to have undocumented earnings.

Workers in the DER file are uniquely identified by a Protected Identification Key (PIK) assigned by Census. The PIK is a confidentiality-protected version of the Social Security Number (SSN). The Census Bureau's Center for Administrative Records Research and Applications (CARRA) matches the DER file to the CPS ASEC. Since the CPS does not currently ask respondents for a SSN, CARRA uses its

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

own record linkage software system, the Person Validation System, to assign a SSN.[8] This assignment relies on a probabilistic matching model based on name, address, date of birth, and gender. The SSN is then converted to a PIK. The SSN from the DER file received from SSA is also converted to a PIK. The CPS ASEC and DER files are matched based on the PIK and do not contain the SSN.

Match rates between the CPS and DER administrative data among earners after 2005 are about 85 percent.[9] The principal regression sample used in our analysis in this draft includes full-time, full-year, non-student wage and salary workers ages 18 to 65 who have positive CPS and DER earnings reported for the prior calendar year. As explained previously, we exclude whole imputations. This 2005-2009 CPS-DER matched regression sample includes 232,939 earners, 128,497 men and 104,442 women, among whom 106,267 men and 85,956 women were earnings respondents in the CPS, implying non-response rates in these CPS samples of 17.3% among men and 17.7% among women (unweighted).

Descriptive data for our data set is provided in Table 1. We focus on measures of earnings and earnings response. Means for all other CPS variables (not shown) are equivalent or nearly equivalent to those seen in public use files. Future versions of the paper will include standard errors and statistical testing of comparisons. Evident in Table 1 is that log wages are higher in the CPS than the DER. The difference in log wages is .06 or about 6 percent for men and about .07 or 7 percent for women. A likely explanation for higher CPS than DER earnings is that DER does not include earnings off the books. The log difference in the case of men substantially overstates the arithmetic percentage difference given the much larger male wage variance in DER versus the CPS (see Blackburn 2007). The higher variance in DER earnings is driven in part by our use of a 2 million dollar cap on DER earnings versus the 1.1 million cap in the internal CPS files (we will subsequently present means using common top-codes). Very high earnings are far more common among men than women. For responding men DER wages are higher ($27.52) than CPS wages ($25.14), but for responding women DER wages ($17.50) are lower than CPS wages ($18.12). For both non-responding men and women CPS wages are higher than DER wages. Comparing CPS wages for respondent and non-respondent men shows respondent men report slightly higher wages than are imputed for non-respondent men ($25.14 vs. $25.09). The same comparison for women shows imputed CPS wages ($18.87) are higher than reported wages ($18.12). The use of proxies and spouse proxies is more prevalent for men than women.

5. Is Earnings Response a Function of Earnings? Non-response across the Distribution

Although evidence is quite limited, previous studies tend to find that there is negative selection into response. That is, as true earnings rise, non-response increases. Testing this is difficult with public use data since we do not observe earnings for those who do not respond. Here we initially follow the approach by Greenlees, et al. (1982), who measure the likelihood of CPS response as a function of matched 1973 administrative (i.e., DER) earnings matched to the CPS, conditional on a rich set of

[8] The Census Bureau changed its consent protocol to match respondents to administrative data during our analysis years. The final year the CPS collected respondent Social Security Number is CPS survey year 2005 (calendar year 2004), the first year of our analysis. In this and prior years respondents provided their SSN and an affirmative agreement allowing a match to administrative data, an "opt-in" consent option. Beginning with survey year 2006 (calendar year 2005), respondents not wanting to be matched to administrative data had to notify the Census Bureau through the website or use a special mail in response, an "opt-out" consent option. If the Census Bureau doesn't receive this notification, the respondent is assigned a SSN using the Person Validation System.

[9] Under the "opt-in" consent option in 2004 the match rate among earners is 61 percent.

covariates. This analysis was done only for white males working full-time/full-year married to a non-working spouse.

Table 2 provides estimates of the relationship between non-response and earnings using our matched CPS-DER sample. It shows the estimation of an equation with earnings non-response as the dependent variable. While the table shows the coefficients of interest (DER log wage and DER wage decile dummies), each equation includes a rich set of covariates.[10] The left panel provides results for men and the right panel for women. In this initial version, we show OLS results; marginal effects using probit estimation (calculated at means of the X's) are highly similar. In contrast to Greenlees et al. (and other prior literature), our coefficients on earnings, for both men and women, point to there being positive rather than negative selection into response, at least on average. That is, the mean tendency across a large representative sample of the workforce is that non-response falls with increases in earnings (a negative coefficient on DER log wage). That said, the coefficient for men, although statistically significant, is very close to zero (-0.011 with s.e. 0.0020), albeit highly significant given our sample size. Among women, we obtain a much larger negative coefficient (-0.0386 with s.e. 0.0024), again indicating that on average non-response declines with earnings, conditional on covariates.

Although this perhaps surprising result provides what we believe are accurate measures of central tendency for these broad samples of men and women, we do not think either this result is particularly informative. Our concern is three-fold. First, the relationship between non-response and earnings may vary over the distribution with differences particularly evident in the tails, thus making measures of central tendency uninformative. Second, the Greenlees et al. result was for a small sample of married white men with non-working spouses, a sample very different from ours. And third, while DER earnings are likely to provide a highly accurate measure of true earnings throughout most of the distribution, the exception is likely to be in the left tail of the distribution, where some share of earnings may be "off the books" and not fully reported to tax authorities. We examine these issues below. Each appears to be important, allowing us to reconcile our results with Greenlees et al. and leading us to draw a more nuanced interpretation of earnings response and non-ignorable response bias.

We first restrict our sample to married white men who are citizens, with spouse present (unlike Greenlees et al., we include those with working spouses). For convenience, we refer to this as our "Mad Men" sample. Note that this sample is likely to have a relatively small proportion of workers in the far left tail of the DER earnings distribution (i.e., those with very low on-the-books earnings). In contrast to the previous coefficient on the log of DER earnings of -0.011 for men, when we estimate this for the more restricted Mad Men sample we obtain a coefficient of 0.04 (not shown in table), highly consistent with Greenlees et al. and other studies concluding that there exists negative selection into response.

Rather than focus on the central tendency, it is more informative to examine how non-response varies across the distribution. Lillard, Smith, and Welch (1986, p. 492) speculated that CPS non-response is likely to be highest in the tails of the distribution (U-shaped), but to the best of our knowledge no study has directly provided such evidence. Since we cannot observe reported CPS earnings for non-respondents, it is difficult to examine this relationship absent matched administrative data on earnings, as is possible with the matched CPS-DER.

[10] The covariates include potential experience, race, marital status, citizenship, education, metropolitan area size, occupation, industry, and year.

To examine nonresponse throughout the distribution, we estimate non-response equations equivalent to those previously shown in Table 2, except that we then group earners into deciles, letting us estimate non-response across the distribution. These coefficients are seen in Table 2. Each coefficient represents the non-response rate at the given DER wage percentile. Readily evident from the coefficients is that there exist U-shaped distributions of non-response, as hypothesized by Lillard et al. (1986). Conditional on the rich set of covariates (wage determinants), non-response among men is particularly high in the 1st decile of the DER wage distribution, about 7½ percentage points higher than in the 2nd decile. Throughout much of the rest of the distribution non-response declines very gradually, but then gradually turns up beginning in the 8th decile. Women exhibit a similar U-shaped pattern of non-response.

Patterns of non-response can be most easily discerned visually (see Figures 1a and 1b, for men and women, respectively). We estimate non-response equations identical to those shown in Table 2, but instead of including the log wage or wage decile dummies, we include dummies for each percentile of the DER wage distribution. The coefficients on these dummies then map out non-response rates throughout the DER wage distribution, conditional on covariates. These rates are shown by the "purple" curve in Figures 1a and 1b (labeled nr_lnwage_der). We next map the relationship between non-response and percentiles of the log wage residual; i.e., the difference between the DER wage and predicted wage (these are shown by "red" squares, labeled nr_ehat). Finally, we map non-response rates with respect to the *predicted* wage from our log wage equation (labeled nr_what). This is seen by the "green" triangles and can be interpreted as showing how non-response varies with workers' earnings attributes (education, demographics, location, etc.) rather than their realized DER wage.

We turn first to the purple curve in Figure 1a, which show male non-response rates with respect to each percentile of the DER wage. The pattern here clearly shows a U-shape, with considerably higher non-response in the lower and upper tails of the distribution, while flat but with a gradual decline from about the 20th to 95th percentiles. Taking these results at face value (which we do, except in the lower tail), suggests that non-response is largely ignorable throughout much of the earnings distribution, varying little with the true level of earnings, conditional on covariates. To the extent that there is a pattern over the 20th to 95th percentiles, it is one consistent with positive selection into response, with non-response gradually declining over much of the distribution, before turning up in the upper tail. Where there most clearly exists a problem is in the far right tail, where we clearly see negative selection into response (those with high realized earnings are less likely to report earnings). Most difficult to interpret is the lower 20% of the DER earnings distribution, where we observe high rates of non-response. We suspect that some of the high non-response in the lowest DER percentiles is due to workers with earnings off the books (thus pushing workers into low percentiles) being reluctant to report earnings in the CPS.

The red curve in Figures 1a shows male non-response rates with respect to percentiles of the wage residual. In the left tail of the distribution (over the first two deciles), where administrative DER earnings are far below predicted earnings, we see very high rates of non-response, thus indicating positive selection in the left tail. In the far right tail of the distribution, where DER earnings are well above predicted earnings, we also see high non-response rates, clearly indicating negative selection into response among very high earners. In short, the evidence clearly indicates that non-response is highest among those who are "strugglers" (under-performers relative to attributes) and "stars" (substantial over-performers).

Finally, the green curve in Figure 1a shows the non-response pattern with respect to men's predicted wage, effectively an index of earnings attributes rather than realized wages. This does not directly measure response bias, but is informative. What we observe is a U-shaped pattern with respect to predicted earnings, but it is less pronounced in the tails of the distribution, particularly the right tail. This pattern suggests that what most affects non-response in the far right tail is the *realization* of very high earnings more so than the attributes associated with high earnings. In the left tail, the rough similarity of all three curves suggests that it is not just low realized earnings that lead to high non-response, but also the individual attributes associated with low earnings.

The evidence for women (Figure 1b) is qualitatively similar to that seen for men, indicating a U-shaped non-response pattern. That said, there are important differences in the magnitudes of the tails. In the lower-end of the wage or wage residual distribution, women exhibit higher rates of non-response than do men. In the right tail of the distribution, women exhibit a minimal increase in non-response, an increase not easily discerned until we move to the highest percentile. In short, calling the female non-response pattern "U-shaped" is a bit of an exaggeration. Of course, women are more likely to have low wages and far less likely to have extremely high wages than are men. Were we to show overlay figures the curves for men and women using a common X-axis (percentiles based on a combined distribution of wages, wage residuals, and predicted wages), we suspect that visual differences in response patterns between men and women would be far less evident. (We have not yet done this.)

We have expressed concern regarding how to interpret results in the left tail of the distribution due to earnings reported in the CPS that are not recorded by DER (our matched sample does not include those with zero DER earnings). In order to examine whether non-response in the left tail results in part from low response among those with earnings off-the books (i.e., not taxed), we can examine the sensitivity of results to inclusion and exclusion of groups most likely to escape or avoid documentation of their earnings. This might include those in household services, some food service occupations, and (among men) construction, as well as for workers foreign born who are not citizens (and possibly those recently arriving in the U.S.). For these groups, we can examine Figures 1a and 1b with these workers removed, which we expect will flatten to some degree non-response in the left tail. Or we can directly examine non-response using data only for these groups of workers and observe how high non-response rates are in the lower tail.[11] Future work will address this.

An alternative way to examine the data is to simply observe the correlates of individual differences between CPS and DER log wages. On average, there is a .07 log wage differential among men and a .06 log differential for women. The higher CPS than DER earnings to some extent results from some earnings being reported in the CPS being off-the-books. We can examine this differential throughout the distribution (of either DER wages or predicted wages), with the expectation that differences are concentrated in the left tail. Relatedly, we will run a regression with the individual log difference of CPS and DER wages as the dependent variable, and then observe how this difference varies

[11] Our CPS and DER wage regressions include sets of industry and occupation dummies, each of which includes a construction dummy. Relative to other industries and occupations, the construction coefficients in the CPS male regression equation indicate higher relative earnings for construction workers than do the coefficients in the DER male regression. These results support the thesis that mean DER earnings across the sample are lower than CPS earnings owing to off-the-book earnings. For a discussion of occupations where off-the-books earnings likely lead to underreporting in DER earnings, see Roemer (2002).

with respect to demographics (e.g., foreign-born, non-citizen), occupation, and industry, among other attributes.

Although preliminary, our conclusion to this point is that if there are severe trouble spots from response bias, these likely occur in the tails of the distribution. In the right tail, high non-response is seen only among the very top 2 percentiles for men and the top percentile for women. But these percentiles correspond roughly to where individual earnings are top coded in the public use CPS. Analysis of earnings in the far right tail is already difficult for researchers using public use files; the added problem of non-response among top earners may inflict little additional damage.[12] In the left tail of the distribution, there are substantial disparities between CPS and DER earnings, but much of this difference may result from high levels of off-the-books earnings. In the next section, we directly examine how wage equation residuals in DER differ between CPS respondents and non-respondents, a straightforward way to evaluate response bias.

6. Examining Patterns of Response Bias: DER Wage Residuals across the Distribution

Perhaps the most direct way to explore patterns of non-ignorable response bias is to compare wage residuals throughout the earnings distribution, with the residuals drawn from DER wage equations including CPS respondents and non-respondents. We first examine summary measures of these differences and then turn to differences across the distribution.

Based on our full-sample DER log wage regression for men, which includes a dense set of covariates, the mean residual for CPS non-respondents is -0.028 and that for CPS respondents is 0.006, a -0.034 difference (by construction, the mean residual for the full sample is zero). This indicates that on average there is weak positive selection into response, with non-respondent men having modestly lower DER earnings than respondents, conditional on measurable covariates. Among women, the pattern of positive selection is somewhat stronger. The mean residual for female CPS non-respondents is -0.071 and that for CPS respondents is 0.015, a -0.086 difference (as compared to -0.034 for men).

Taken at face value (i.e., regarding DER as a measure of true earnings), these magnitudes are small but non-trivial. Based on the 17.3% non-response rate in our male sample, the overall upward bias in CPS earnings due to positive selection into response would be about half of a percent (.173 times -0.034 equals -0.006). For women, the bias is a more substantial one and half percent (.177 times -0.086 equals -0.015). Taken together, this would imply that the gender wage gap is understated by about 1-2 percentage points due to response bias.

As previously discussed, we are reluctant to accept these results at face value, given concerns that some of the residual differences reflect actual earnings reported in the CPS but not in DER administrative records. Moreover, we suspect that such differences are most evident in the lower tail of the distribution where non-response is particularly high and off-the-books earnings may be most prevalent. To examine this, Table 3 shows the mean of residuals for the respondents and non-respondents at selected percentiles of the overall DER wage distribution, separately for men and women. Also shown is the difference in

[12] Researchers using the CPS often assign mean earnings above the top-code based on information provided by Census or by researchers using protected files (Larrimore et al. 2008). Because very high earners are less likely to report earnings in the CPS, there will be some understatement of high-end earnings due to non-ignorable response bias.

residuals between non-respondents and respondents at each percentile. Two stylized facts are readily evident in Table 3. First, severe response bias (i.e., large residuals) is concentrated in the lowest percentiles of the earnings distribution, the very part of the distribution where we suspect underreporting DER earnings. Second, we also see response bias in the very top of the distribution, but in this case negative rather than positive residuals. As previously concluded, in the lower tail of the distribution there is positive selection into response whereas in the upper tail there is negative selection into response.[13]

How much of the apparent response bias is due to earnings off-the-books will be examined in a subsequent version in which we vary our samples focusing separately on demographic characteristics and industry and occupation groups most likely and least likely to have earnings underreported in DER. If such differences are substantial, this is mixed news. Such news would be "troubling" in that substantial off-the-books earnings not reported in DER make it more difficult to use DER to measure the degree of response bias. Such news would be "encouraging" in that it implies that CPS earnings reports capture a good portion of such earnings and that non-ignorable response bias in the left tail of the distribution may be far less than suggested by our evidence. What is clear from our evidence to date is that over most of the wage distribution, non-ignorable response bias appears to be quite limited. It is clearly a second-order concern compared to the first order "match bias" that can arise from using Census imputed earnings (Bollinger and Hirsch 2006).

7. Non-response Implications for Earnings Regression Coefficient Estimates

Bollinger and Hirsch (2013) examine how non-ignorable response bias may affect CPS earnings regression coefficient estimates. Based on models that attempt to account for selection into response using both CPS-ASEC and CPS-ORG files, they compare earnings regression coefficients from their full-sample selection models with those from OLS (without a selection term) based on the sample of respondents (i.e., imputed earners are omitted). They conclude that for both men and women, differences due to response bias show up primarily in the intercepts and that earnings slope coefficients are nearly identical in the selection model and standard OLS. (Owing to imputation match bias, OLS slope coefficients are substantially different using full samples that include allocated earners.)

We can see if this encouraging conclusion holds up based on evidence from our validation sample. Using the DER wage sample, which provides earnings for CPS non-respondents, we can estimate separate wage equations for the CPS respondents and CPS non-respondents. A comparison of intercepts and slope coefficients allow us to see if the Bollinger-Hirsch conclusion holds up. We would not be surprised to see coefficient differences on selected demographic characteristics (e.g., foreign born noncitizen) and industries and occupations where off-the-books earnings are most likely. Such differences need not indicate that there is a problem from response bias for researchers using CPS data to estimate slope coefficients, but would make it more difficult to clearly rule out response bias as a concern.

We expect to conduct such analysis shortly.

[13] For both respondents and non-respondents, residuals are mechanically negative (positive) in the left (right) tails of the distribution. Our conclusions are based on *differences* in residuals for respondents and non-respondents throughout the distribution.

8. How and Why Proxy-reported CPS Earnings Differ from Self Reports

Roughly half of all earnings reports in the CPS are from proxies. Using the matched CPS-ASEC/DER data we have examined how use of proxies affects earnings reports. However, this analysis has not undergone Census disclosure review and cannot be reported in this draft. We will first briefly summarize below prior unpublished research (Bollinger and Hirsch 2009) on proxies using public use versions of the CPS-ORG and CPS-ASEC files. We will then describe how we are analyzing proxy effects using the matched CPS-administrative data.

Bollinger and Hirsch (2009) examine the wage effects of proxy use based on cross-section and panel CPS data. Cross section analysis simply identifies the partial correlation of proxy reports with earnings (conditional on a dense set of wage determinants), but cannot reliably distinguish between earnings differences due to proxy misreporting versus earnings differences due to unobserved worker heterogeneity correlated with a wage earner having a proxy report. Short (one-year) CPS panels are nicely suited for such an analysis, accounting for worker fixed effects and identifying proxy reporting effects based on workers who self-report in one year and have a proxy report the next year (or vice-versa). Bollinger and Hirsch also introduce the distinction between spouse proxies and non-spouse proxies.[14]

Bollinger and Hirsch (2009) find similar patterns in the ORGs and ASEC. With cross-section analysis, they find that spouse reports are nearly identical to self-reports, while non-spouse reports are substantially lower than self or spouse reports. However, when they move to panel analyses, which arguably approximates proxy misreporting, they find that the effects of spouse and non-spouse proxies on earnings reports are similar, being one to three percent lower than self-reports (wives tend to understate husband earnings, while evidence for the reverse is weak). Bollinger and Hirsch conclude that in general proxy reports are quite close to self-reports, a conclusion similar to that seen in earlier validation studies (e.g., Mellow and Sider 1983). However, the large negative coefficients for non-spouse proxies found in the cross section (and near zero coefficients for spouses) suggest that these proxy coefficients are capturing unmeasured worker productivity effects.

Using the matched CPS-ASEC/DER sample, we can observe whether administrative earnings in DER, where there are no proxies, vary with respect to proxy use in the CPS. That is, we simply include in the DER equation "phantom" dummies for use of a spouse and non-spouse proxy in the CPS. What we find is that coefficients on the phantom dummies are just slightly smaller (in absolute value) than are the coefficients seen in the CPS. What these DER coefficients must reflect are worker fixed effects (unobserved heterogeneity) that shows up in both the DER and CPS earnings. Thus, we tentatively conclude that proxy reports in the CPS are on average reasonably accurate. Although proxy reports (in particular those from non-spouse proxies) are associated with wage differences in the CPS, these mostly reflect actual earnings differences and not misreporting of earnings.

[14] Reynolds and Wenger (2012) use CPS-ORG panels over a long time period to examine the effect of proxies, focusing on gender differences and changes in proxy patterns over time. They do distinguish between spouse and non-spouse proxies.

9. Conclusion

This paper has addressed three questions not adequately examined in prior literature. First, we address the question of how non-response varies across the earnings distribution, a difficult question to answer absent information on non-respondents' earnings. Our preliminary findings include the following. Non-response across the earnings distribution, conditional on covariates, is U-shaped, with left-tail "strugglers" and right-tail "stars" being least likely to report earnings. Women have particularly high non-response in the left tail; men have high non-response in the far right tail. Second, we ask whether response bias is ignorable; that is, whether respondents and non-respondents have equivalent earnings, conditional on covariates, throughout the earnings distribution. Throughout much of the distribution there is little correlation between response and earnings, implying that non-response is largely ignorable over this range, but with possible trouble in the tails. And third, we examine whether proxy responses, which account for half of all CPS earnings reports, are reliable. We find that proxy response is correlated with earnings, conditioning on covariates, but this largely reflects unmeasured worker heterogeneity and not misreporting of earnings.

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*, forthcoming.

Blackburn, McKinley L. 2007. "Estimating Wage Differentials without Logarithms." *Labour Economics* 14:1 (2007): 73-98.

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, "Wage Gap Estimates with Proxies and Nonresponse," Unpublished manuscript, November 2009.

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

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.

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.

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.

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.

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

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

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.

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

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

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

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

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.

Table 1: Selected Summary Statistics for CPS-ASEC/DER Estimation Sample

Men			Women		
Variable	**Mean**	**Std. Dev.**	**Variable**	**Mean**	**Std. Dev.**
Wage-CPS	25.10	25.65	Wage-CPS	18.74	16.68
Wage-DER	25.36	72.74	Wage-DER	17.90	20.14
lnW-CPS	3.003	0.65	lnW-CPS	2.746	0.60
lnW-DER	2.944	0.80	lnW-DER	2.676	0.69
CPS Non-respondents (Earnings Imputation Rate)	0.173	0.38	CPS Non-respondents (Earnings Imputation Rate)	0.177	0.38
Wage-DER (CPS Respondents)	27.52	127.62	Wage-DER (CPS Respondents)	17.50	35.74
Wage-DER (CPS Non-respondents)	24.91	54.63	Wage-DER (CPS Non-respondents)	17.98	14.77
Wage-CPS (CPS Respondents)	25.14	29.93	Wage-CPS (CPS Respondents)	18.12	17.07
Wage-CPS (CPS Non-Respondents)	25.09	24.66	Wage-CPS (CPS Non-Respondents)	18.87	16.59
CPS proxies	0.550	0.50	CPS proxies	0.406	0.49
Spouse proxies	0.401	0.49	Spouse proxies	0.258	0.44
Nonspouse proxies	0.148	0.36	Nonspouse proxies	0.148	0.36
Observations	128,497		Observations	104,442	

Sources: U.S. Census Bureau, Current Population Survey, 2005-2009 Annual Social and Economic Supplement. For information on sampling and nonsampling error, see www.census.gov/apsd/techdoc/cps/cpsmar13.pdf.

Social Security Administration, Detailed Earnings Record, 2004-2008.

Table 2: DER Log Wage and Percentile Coefficients in Non-Response Equation

	Men		Women	
Variable	**coeff.**	**s.e.**	**coeff.**	**s.e.**
DER LogWage	-0.011	0.0020	-0.039	0.0024
R-sq	0.017		0.019	
Observations	128,497		104,442	
DER Wage Decile				
1st	0.069	0.0082	0.097	0.0107
2nd	-0.007	0.0080	0.021	0.0106
3rd	-0.011	0.0081	0.002	0.0106
4th	-0.026	0.0081	-0.002	0.0106
5th	-0.028	0.0081	-0.004	0.0107
6th	-0.027	0.0082	-0.006	0.0107
7th	-0.033	0.0082	-0.016	0.0107
8th	-0.020	0.0082	-0.011	0.0108
9th	-0.012	0.0083	-0.014	0.0109
10th	0.017	0.0085	0.006	0.0111
R-sq	0.191		0.194	
Observations	131,084		105,652	

Ordinary Least Squares estimation with robust standard errors. Specifications include controls for potential experience, race, marital status, citizenship, education, metropolitan area size, occupation, industry, and year.

Sources: U.S. Census Bureau, Current Population Survey, 2005-2009 Annual Social and Economic Supplement. For information on sampling and nonsampling error, see www.census.gov/apsd/techdoc/cps/cpsmar13.pdf.

Social Security Administration, Detailed Earnings Record, 2004-2008.

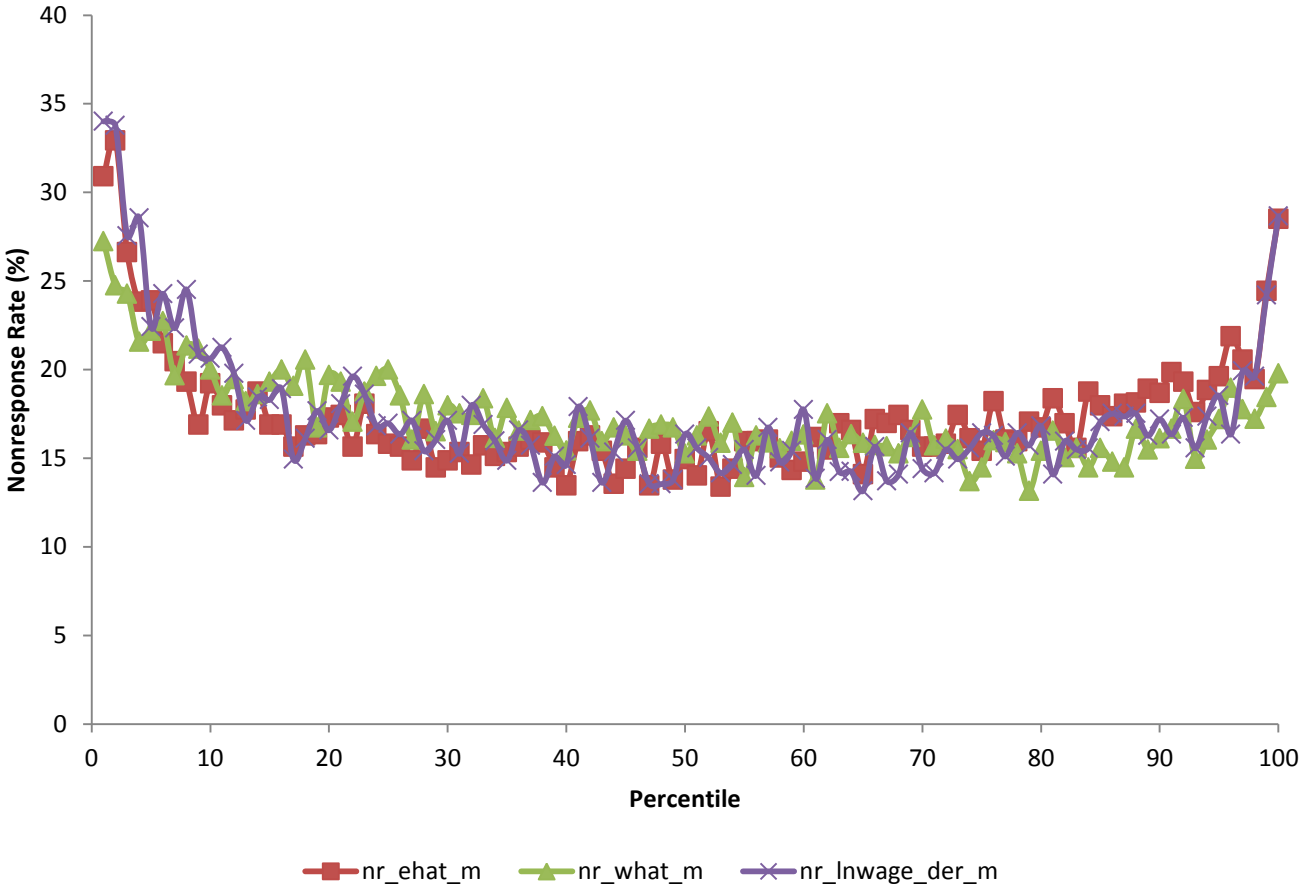
Table 3: Differences in DER lnWage Residuals for CPS Non-Respondents and CPS Respondents

	Men			Women		
DER Wage Percentile	Non-Responders	Responders	NR-R Difference	Non-Responders	Responders	NR-R Difference
1%	-2.748	-1.917	**-0.831**	-2.750	-1.708	**-1.042**
2%	-1.212	-0.829	**-0.383**	-1.199	-0.756	**-0.443**
10%	-0.754	-0.567	**-0.187**	-0.741	-0.514	**-0.227**
25%	-0.319	-0.253	**-0.066**	-0.316	-0.223	**-0.093**
50%	0.041	0.041	**0.000**	0.018	0.049	**-0.031**
75%	0.358	0.319	**0.039**	0.312	0.312	**0.000**
90%	0.655	0.586	**0.069**	0.586	0.563	**0.023**
95%	0.864	0.763	**0.101**	0.771	0.726	**0.045**
99%	1.500	1.239	**0.261**	1.200	1.088	**0.112**
Mean	-0.028	0.006	-0.034	-0.071	0.015	-0.086
Std Dev	0.835	0.618	0.217	0.699	0.538	0.161
Variance	0.696	0.381	0.315	0.489	0.289	0.200
Obs	22,274	106,223		18,477	85,965	

Sources: U.S. Census Bureau, Current Population Survey, 2005-2009 Annual Social and Economic Supplement. For information on sampling and nonsampling error, see www.census.gov/apsd/techdoc/cps/cpsmar13.pdf.

Social Security Administration, Detailed Earnings Record, 2004-2008.

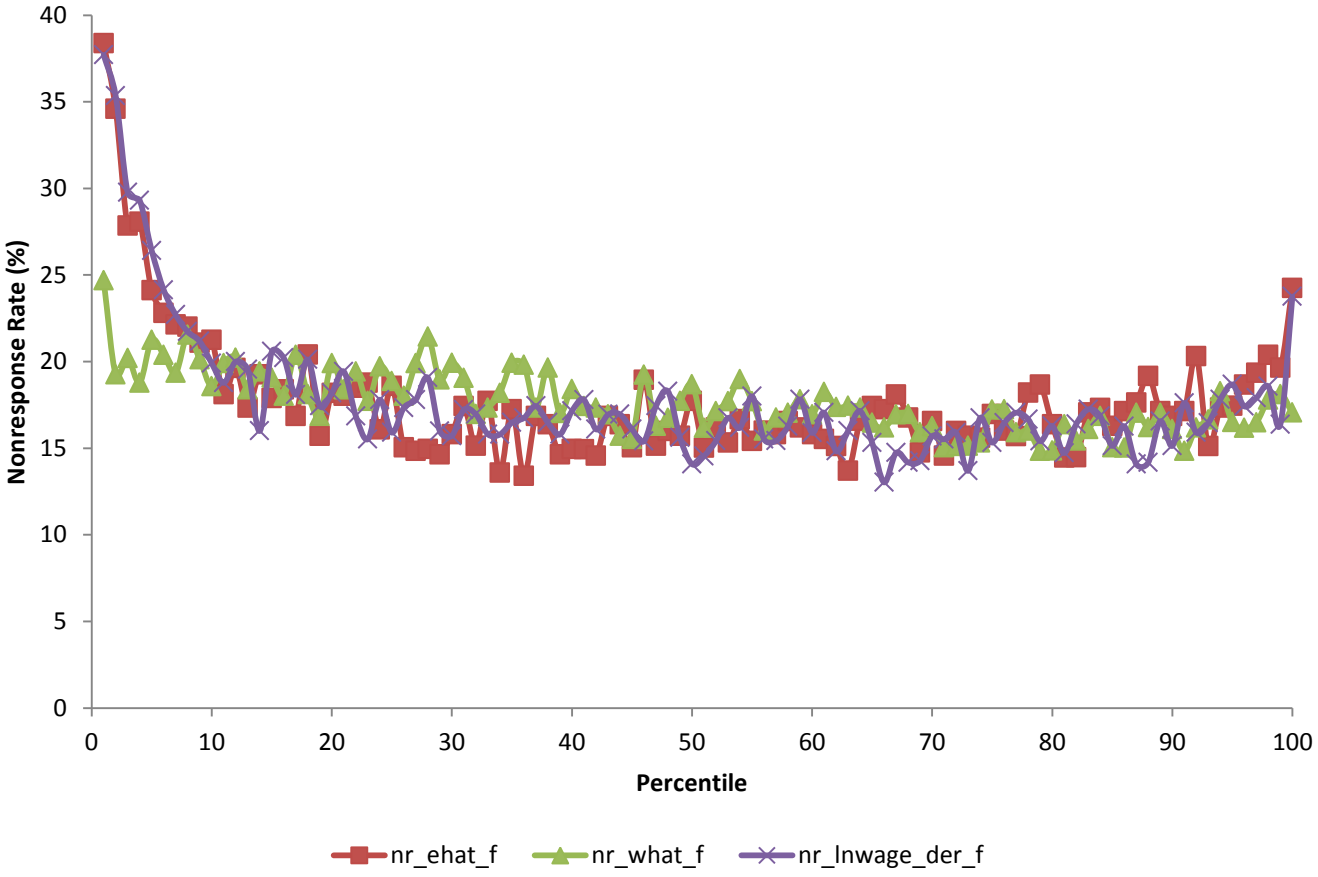
Figure 1a: CPS Nonresponse by Percentiles of Residual, Predicted, & Actual DER Wages, Men



Sources: U.S. Census Bureau, Current Population Survey, 2005-2009 Annual Social and Economic Supplement. For information on sampling and nonsampling error, see www.census.gov/apsd/techdoc/cps/cpsmar13.pdf.

Social Security Administration, Detailed Earnings Record, 2004-2008.

Figure 1b: CPS Nonresponse by Percentiles of Residual, Predicted, & Actual DER Wages, Women



Sources: U.S. Census Bureau, Current Population Survey, 2005-2009 Annual Social and Economic Supplement. For information on sampling and nonsampling error, see www.census.gov/apsd/techdoc/cps/cpsmar13.pdf.

Social Security Administration, Detailed Earnings Record, 2004-2008.