Examining the Black-White Earnings Differential with Administrative Records

Michael Gideon

Misty L. Heggeness[ψ]

Marta Murray-Close

U.S. Census Bureau

Samuel L. Myers, Jr.

University of Minnesota

December 22, 2016

SEHSD Working Paper 2017-32

[ψ] Corresponding author; comments or inquiries can be sent to Misty Heggeness at misty.l.heggeness@census.gov.

Abstract

In this paper, we compare estimates of the black-white male earnings gap using self-reported earnings from the Census Bureau's Current Population Survey (CPS) Annual Social and Economic Supplement (ASEC) and reported earnings from the Social Security Administration's Detailed Earnings (DER) File. We find that using self-reported earnings data from the CPS masks the magnitude of inequality between black and white earnings and that research using imputed earnings with self-reported data has the potential to skew earnings gap estimates generated from self-reported household survey data compared to administrative records. Overall, our results provide informative and relevant information for understanding the extent to which self-reported earnings data and reported earnings from administrative records influences the estimation of the black-white earnings gap. The results will also help inform the potential for improvements to official national inequality estimates and measurements.

Introduction

Economic research on earnings and income relies primarily on self-reported earnings from household surveys like the Current Population Survey (CPS).[1] One area, which has relied primarily, if not solely, on self-reported earnings, is in the study of wage and earnings inequality. Federal agencies calculate Gini coefficients and other inequality measures using the CPS and other household surveys (Proctor, Semega, and Kollar 2016; Posey 2016). Other investigative studies as early as the 1970s have focused on analyzing black-white earnings gaps with the CPS and other household surveys (Masters 1974; Smith and Welch 1977; Darity et al. 1998; O'Neill 1990; Arias et al. 2004; Gabriel 2004; Cunningham and Jacobsen 2008).

In this paper, we examine the effect, if any, of using self-reported household survey earnings data on inequality, specifically the black-white earnings gap. Our study is twofold. First, we examine whether differences in earnings inequality exist between the CPS and administrative tax records. To the extent that individuals self-reported earnings inaccurately, with less precision, or to the extent that top coding influences inequality measures, we capture the overall magnitude (if any) of these effects. Second, we examine the role, if any, that imputations play on self-reported earnings data within household survey data. Overall, this paper contributes to the literature by expanding our understanding of potential issues to be aware of when using household survey earnings data to study and measure inequality, particularly differences in earnings across demographic groups.

[1] For more information on the Current Population Survey, see: https://www.census.gov/programs-surveys/cps.html or https://cps.ipums.org/cps/.

Background

For years, experts have studied the reasons why a black-white earnings gap is so persistent within the United States (Masters 1974; Smith and Welch 1977; Darity et al. 1998; O'Neill 1990; Arias et al. 2004; Gabriel 2004; Cunningham and Jacobsen 2008). The economic literature on earnings differentials emphasizes understanding the difference between labor market discrimination and differences due to age, educational attainment, level of skill, and other geographic and demographic constraints that differentiate groups and confound our understanding of the earnings gap. In order to understand true discrimination in the labor market, this literature highlights the need to have quality data on the other aspects driving differences in earnings gaps and to control for those differences. There is very little discussion in this literature of the quality of earnings data and the potential for erroneous estimation of inequality due to confounding issues in data collection and reporting.

Recent studies have used administrative records to study earnings non-response in surveys (Abowd and Stinson 2013; Hokayem et al. 2015, 2016; Bollinger et al. 2015). These studies have found that imputed earnings match relatively well with administrative records for most of the earnings distribution. However, at each tail end of the distribution, there are problems. One particular problem arises from imputations for individuals with low earnings. These earnings tend to be lower than those reported in administrative records. A second problem arises from imputations for high earners whose self-reported earnings tend to be higher than those reported in administrative records. These results could affect inequality measures. In this paper, we examine the extent to which using administrative records on earnings changes our estimation of the black-white earnings gap and our understanding of inequality in earnings between black and white men.

Data and Methods

Data are from the Current Population Survey (CPS) Annual Social and Economic Supplement (ASEC) and the Social Security Administration's Detailed Earnings File (DER). Each year, between February and April, the CPS surveys approximately 99,000 households about their incomes and characteristics of their employment in the previous calendar year. We use demographic and socioeconomic variables from the CPS, a measure of earnings using the CPS, and a measure of earnings using the DER. We use data from CPS for 2005-2013, corresponding to tax years (and DER earnings data for) 2004-2012. The years are limited on the front end due to geography and concerns about the linkage to the administrative records, and on the back end due to availability of tax earnings data.

For this project, we use information from the CPS on all wage and salary earnings, and the characteristics of the longest-held job, particularly information used to determine full-time full-year work status and the class of employer (private, public, or self-employed). We include all wage and salary earnings because CPS does not collect information about every job separately. Administrative earnings data come from the DER.[2] We define annual earnings for each job in the DER as the sum of Box 1 (wages, salaries, bonuses, etc.) and Box 12 amounts (tax-deferred contributions to employer-sponsored retirement plans, such as 401(k)s). Annual earnings for each job are summed together to get yearly earnings. Both earnings variables are adjusted to 2015 dollars using the Consumer Price Index (CPI).

[2] The DER contains information on all jobs with W-2 forms, as well as self-employment information from tax returns. We focus on private- and public-sector workers, and therefore focus on earnings in the DER that come from W-2 records.

Sample restrictions

The baseline sample consists of men who were between the ages of 25 and 64 at the time of their CPS interview, who identified their race and ethnicity as non-Hispanic white or non-Hispanic black, and whose longest-held job was in a non-military occupation.[3] To minimize the impact of racial differences in work hours on our estimates of differences in earnings, we further restrict the sample to full-time full-year workers.[4] To maximize the comparability of our CPS earnings measure with our DER earnings measure, which does not include earnings from self-employment, we restrict the sample to men whose longest-held job was in the private or public sector. Finally, we restrict the sample to individuals with strictly positive earnings in both the CPS and the DER in a given year.

By restricting the baseline sample to full-time full-year workers in the CPS, all observations have strictly positive earnings in the CPS, and we do not have cases in which people report zero earnings in CPS but have earnings in the DER. However, there are two reasons a person in the CPS might not match to the DER. Most importantly, he or she might not have been assigned a unique personal identification key (PIK). The PIK is a unique identifier developed by the Census Bureau that facilitates linking individuals to other individual-level datasets.[5] Respondents who do not have validated PIKs are removed and observations are re-weighted to account for the probability of being assigned a PIK. Second, people can have a validated PIK, report working in the CPS, but not have earnings reports in the DER. This

[3] Less than 1.5 percent of the baseline sample reported multiple racial identifications in the CPS. These people were assigned to a single-race category using a standard recoded race and ethnicity variable. When revising this work, we plan to restrict our estimation samples to people who reported only one racial identification.
[4] Full-time full-year employment is defined as working at least 50 weeks during the year and usually working at least 35 hours per week during the weeks worked.
[5] See Wagner and Layne (2014) for more information on the PIK assignment process.

problem arises when there are informal jobs that do not generate W-2 forms. This is uncommon after making the other sample restrictions and removing imputations.[6]

Imputations

The baseline sample, as shown in Table 1, includes imputed data. We also estimate the models using a restricted sample that excludes imputations. As highlighted by Meyer et al. (2015), survey non-response rates have increased substantially over the past couple decades. Hokayem et al. (2015) show that nonresponse affects measures of poverty, suggesting that imputations could bias results. More generally, nonresponse rates are highest in the tails of the income distribution (Bollinger et al., 2015). Because blacks and whites have different earnings distributions and imputation rates, nonresponse (and errors due to imputation) could explain differences in the black-white earnings gap when using earnings measures from CPS versus the DER. We therefore explore the role of imputations in driving differences in earnings gaps using CPS and DER.[7]

[6] If informal arrangements vary across time differently by race, then changes in the difference in the black-white earnings gaps using CPS and DER over time could be impacted by the changes in the samples.
[7] While respondents occasionally choose not to answer questions about age, race, ethnicity, or education level, it is much more common for respondents to omit answers to questions about hours worked or earnings.

Table 1. Sample of full-time year-round male workers aged 25 to 64 working in the public or private sector by imputation status

	Sample with Imputations		Sample without Imputations		(%) Imputed	
	Black	White	Black	White	Black	White
2005	2,965	24,585	1,768	17,753	40.4	27.8
2006	2,870	24,449	1,862	18,360	35.1	24.9
2007	3,025	24,220	1,913	17,739	36.8	26.8
2008	3,091	23,947	1,963	17,731	36.5	26.0
2009	2,924	23,363	1,954	17,526	33.2	25.0
2010	2,894	21,916	1,915	15,836	33.8	27.7
2011	2,692	21,567	1,682	15,075	37.5	30.1
2012	2,674	21,440	1,682	15,198	37.1	29.1
2013	2,705	21,412	1,697	14,932	37.3	30.3

Source: Current Population Survey (CPS) Annual Social and Economic Characteristics (ASEC) Files linked to the Detailed Earnings Record (DER), 2005-2013
Note: White refers to White, non-Hispanic and Black refers to Black, non-Hispanic

Respondents may refuse to answer specific questions in the CPS or refuse to answer the entire Annual Social and Economic Supplement survey. To explore the role of nonresponse and imputations, we remove observations that are imputed due to supplement nonresponse and observations that have imputed values for earnings from their longest held job.[8]

Estimates computed on this restricted sample are less vulnerable than estimates computed on the baseline sample to bias from data-processing decisions. There is, however, a cost to excluding the imputations. Unless the underlying data are missing at random, observations with full-record or earnings imputations are likely to differ from those without imputations. The analyses in this paper do not adjust the sample weights to account for the impact of excluding the imputations on the composition of our estimation sample. Consequently, estimates computed on

[8] It's possible that people report their earnings from their main job but have imputed values for their other earnings. These cases are treated as not being imputed. Any differences we find due to removing imputations are presumably understated, since we would remove additional observations in these other cases.

the restricted sample may not accurately represent the target population of workers. In future work, we will investigate methods to adjust weights to generate population-level estimates.

Descriptive results

Figure 1 shows the ratio of black to white mean male earnings as measured by the CPS and the DER, along with 90-percent confidence intervals, for the baseline sample described above, over the CPS survey years 2005 to 2013. To obtain nationally representative results, point estimates were computed using the CPS person weights. Standard errors and corresponding confidence intervals were computed using the CPS replicate weights. Both the person weights and the replicate weights were adjusted for the probability that a CPS respondent was successfully matched with an administrative earnings record.

Figure 1 has two notable features. First, the ratio of black to white earnings was relatively stable over the study period, although the point estimates are slightly higher at the end of the period than the beginning. Second, black men had higher earnings relative to white men in the CPS than the DER. Specifically, the ratio of black to white mean male earnings ranges from 0.63 to 0.72 in the CPS but just 0.58 to 0.64 in the DER. Figure 1 clearly shows that the black to white earnings ratio is higher in the CPS data than the DER data for all years. With the exception of the first and last years, the two series have non-overlapping 90-percent confidence intervals. To the extent that that earnings data from administrative records, which are not subject to recall or rounding errors or to top coding, are more accurate than self-reported earnings, these results suggest that the CPS data underestimate the gap in black-white earnings.

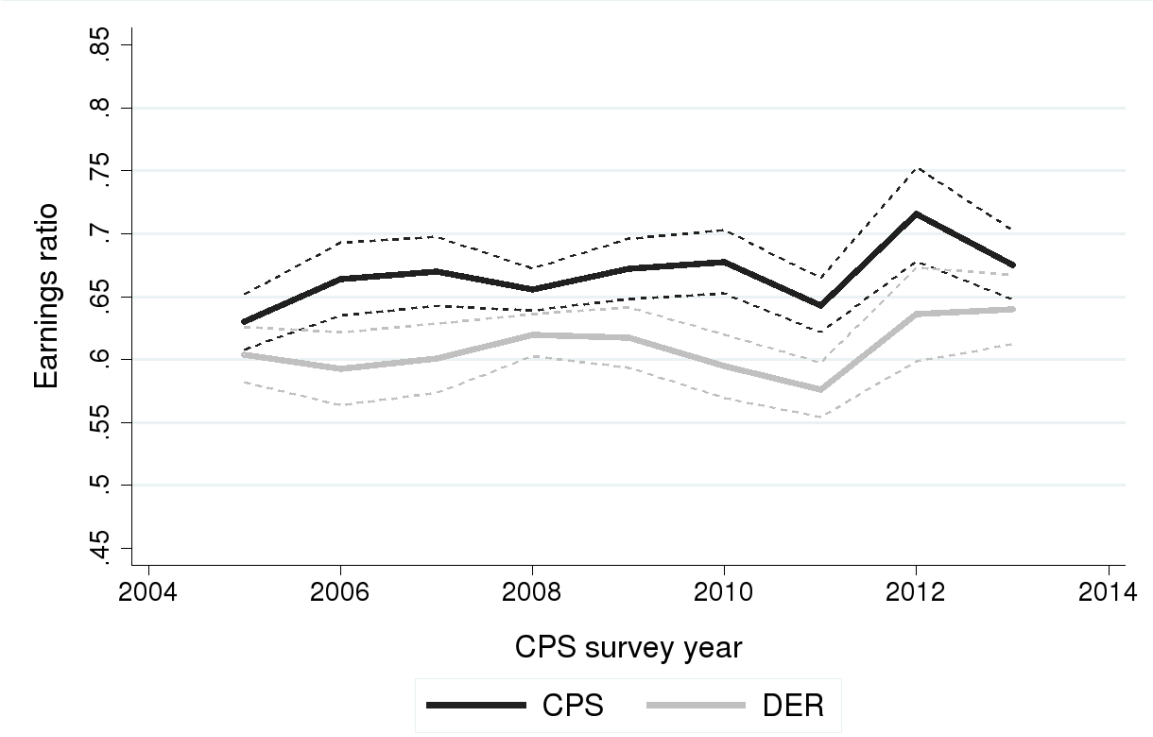

Figure 1
Ratio of black to white mean male earnings, baseline sample

Notes: Data are CPS Annual Social and Economic Supplement survey responses linked to Detailed Earnings Records from the Social Security Administration. Dashed lines are upper and lower bounds of 90-percent confidence intervals. Point estimates were computed using CPS person weights adjusted for the probability of having positive DER earnings. Confidence intervals were computed using CPS replicate weights. Baseline sample is black and white non-Hispanic men age 25 to 64 who were employed full-time full-year in a non-military occupation and had positive earnings in both the CPS and the DER.

Figure 2 shows the ratio of black to white mean male earnings for the restricted sample that excludes observations with full-record or earnings imputations in the CPS. Excluding observations with imputations reduces the difference between estimates of relative black earnings based on self-reports from the CPS and estimates based on administrative records from the DER. This convergence in the CPS and DER estimates reflects both a decrease in the CPS earnings ratios and an increase in the DER earnings ratios. The change in the CPS estimates may

be due to a reduction in bias from data processing decisions, shifts in the composition of the estimation sample, or both. In contrast, because the DER earnings data are not affected by the CPS imputations, the change in the DER estimates must be due to shifts in the composition of the estimation sample.

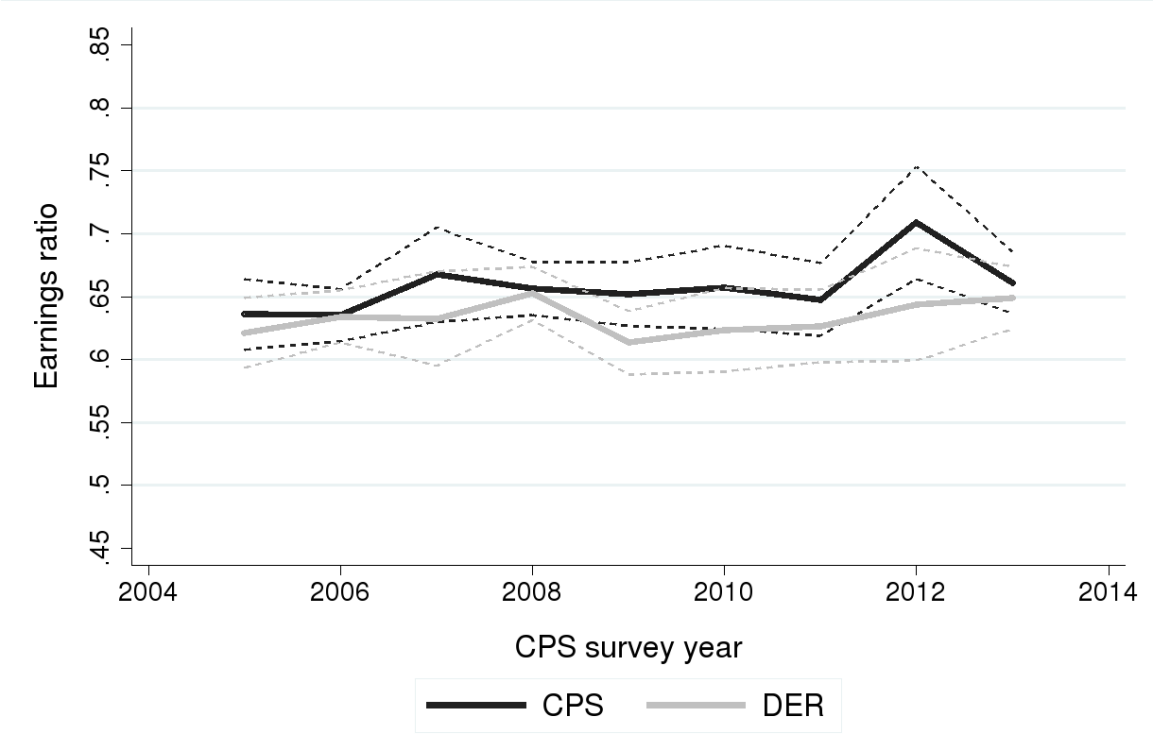

Figure 2
Ratio of black to white mean male earnings, restricted sample

Notes: Data are CPS Annual Social and Economic Supplement survey responses linked to Detailed Earnings Records from the Social Security Administration. Dashed lines are upper and lower bounds of 90-percent confidence intervals. Point estimates were computed using CPS person weights adjusted for the probability of having positive DER earnings. Confidence intervals were computed using CPS replicate weights. Restricted sample is baseline sample, as described in the notes to Figure 1, excluding observations with full-record or earnings imputations in the CPS.

The Effects of Imputation and Source of Data on Explained Black-White Gaps in Earnings

As we have seen in the descriptive analysis, measured black-white gaps in mean earnings often are larger using the DER data than using the CPS data. Figure 3 provides a recapitulation of this key finding, but using differences in the mean of log earnings. The figure focuses on white non-Hispanic males versus black non-Hispanic males, ages 25-64, full-time, year-round workers with positive earnings. The sample includes persons who worked in the public and private sectors, but excludes those whose only earnings are self-employed earnings or whose primary job is self-employment in the CPS. The sample is drawn from the civilian, non-institutionalized population for years 2006, 2009, 2010 and 2012. We chose these years primarily because they fall before, during, lagged, and after the Great Recession and were interested in understanding any differences between these particular periods.

In reporting the white-black difference in log-earnings, Figure 3 shows *larger gaps* in earnings for higher numbers and *smaller gaps* in earnings for lower numbers. This is understood to translate to higher ratios of black to white earnings when there are smaller gaps in earnings and smaller ratios of black to white earnings when there are larger gaps in earnings. Thus, Figure 3 reflects the core findings in Figure 1 showing that the CPS measures of black-white earnings gaps are smaller than the DER earnings gaps.

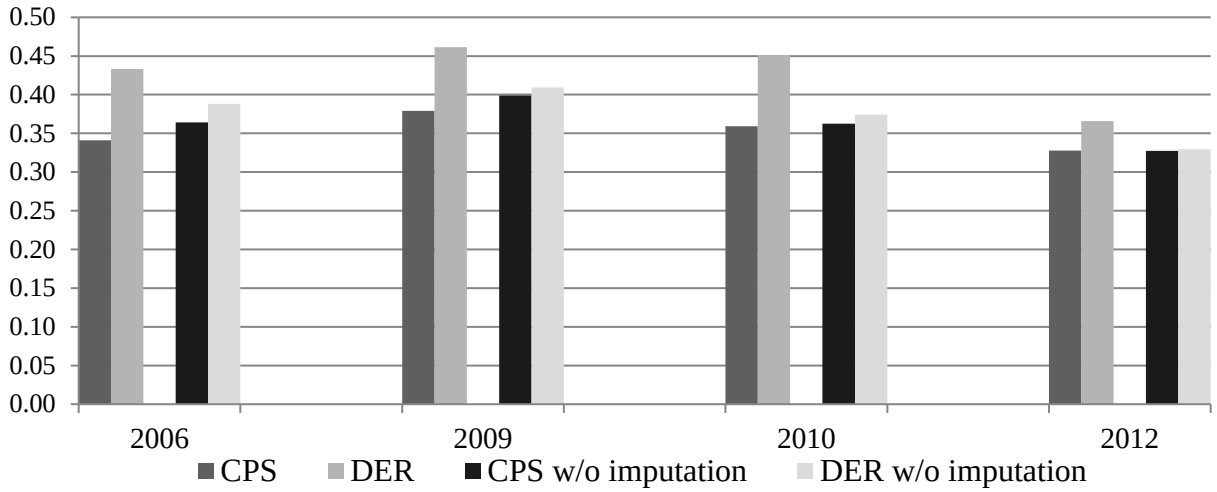

Figure 3
Difference in Log-Earnings – CPS versus DER, Various Years

In each year with imputation the difference in the log-earnings between whites and blacks is always smaller using the CPS data than using the DER data. Without imputation, the gaps are slightly larger in the DER data than CPS data for 2006 but barely different for the remaining years.

One would want to know, however, whether the source of data affects the results of the decomposition of the gaps into explained versus unexplained portions. Ignoring for a moment the impacts of imputation on the decomposition, we estimate the following basic Mincer earnings equations for blacks and whites (denoted by the superscript k) for the years t = 2006, 2009, 2010 and 2012:

$$\ln(y_{it}^{k}) = \sum \beta_{it}^{k} x_{it}^{k} + \varepsilon_{it}^{k}$$

Independent variables include: age, age-squared, highest degree (e.g. high school diploma), and a dummy variable for public- versus private-sector employment. The overall gap in average

earnings between blacks and whites, $\overline{\ln(y)}_t^W - \overline{\ln(y)}_t^B$, can be decomposed using a conventional Oaxaca-Blinder method into an explained portion, unexplained portion, and an interaction:

$$\overline{\ln(y)}_t^W - \overline{\ln(y)}_t^B = \sum \beta_t^B(\bar{x}_t^W - \bar{x}_t^B) + \sum(\beta_t^W - \beta_t^B)\bar{x}_t^B + \sum(\beta_t^W - \beta_t^B)(\bar{x}_t^W - \bar{x}_t^B)$$

We report in Table 2 the results of the full decomposition for each year showing the percentage of the log-earnings gap attributable to: endowments, coefficients, and interaction. The interpretation of the endowment effect is the percentage of the racial gap in log-earnings explained by the independent variables – education, age, and sector of employment. The interpretation of the coefficient effect is the difference in earnings attributable to differences in treatment or the pure discrimination effect. The remainder is an interaction between the treatment effects and the endowment effects. Table 2 reveals that the endowment effect – the portion of the gap explained by age, education and sector of employment – is larger in the CPS data than the DER data for each year. The differential treatment effects correspondingly are larger in the DER data than in the CPS data for each year. Based on these results, to the extent that the DER data are more accurate than the CPS data, one can conclude that using the CPS data *overestimates* explained portions of the black-white log-earnings gaps and *underestimates* measures of discrimination. This is true across different periods: before the Great Recession (2006); during the recession (2009/2010 lagged recession effects); and post-recession (2012).

Table 2. Decomposition of Log-Earnings Gaps for CPS versus DER Data with Imputations

	2006	2009	2010	2012
Endowments				
CPS	24%	26%	25%	20%
DER	19%	24%	22%	18%
Coefficients				
CPS	69%	68%	66%	73%
DER	73%	72%	71%	75%
Interaction				
CPS	8%	6%	8%	6%
DER	7%	4%	7%	7%

Effects of Imputation

As described above, individuals who are unwilling or unable to report earnings have the amounts "allocated" (or "imputed") using hot deck estimation procedures. About 8 to 14 percent of the sample each year does not answer earnings questions on the March supplement. The Census Bureau uses hot deck methods to impute responses when missing data exists, and cases where responses are imputed are flagged within the data. Omitting the cases without earnings produces a sample that might not be representative of the entire population but overcomes the problem of whether there is a bias introduced in the measures of discrimination by using imputed values.

Table 3 reports the results of the measured discrimination comparing the coefficient effects between the sample with imputed values versus the sample without imputed values. The table reports these results for each year and for CPS versus DER. The measured discrimination is almost exactly the same for the CPS and DER data when imputed values are excluded. The percentage of the log-earnings gap due to differences in coefficients ranges from 71 percent in

2006 and 2009 to 74 percent in 2010 in both the CPS and DER data. The discrimination measure in 2012 is 73 percent in the CPS data but 72 percent in the DER data.

Table 3. Comparison of Measured Discrimination with and without Imputations, Various Years

	2006	2009	2010	2012
CPS Coefficients				
with Imputations	69%	68%	66%	73%
without Imputations	71%	71%	74%	73%
DER Coefficients				
with Imputations	73%	72%	71%	75%
without Imputations	71%	71%	74%	72%

Measured discrimination in the CPS data with imputation is generally smaller with imputed values than without imputed values. The percentage of the gap in 2006 attributable to differences in coefficients is 69 versus 71 percent with versus without imputation; the percentage of the gap in 2009 attributable to differences in coefficients is 68 versus 71 percent with versus without imputation; the percentage of the gap in 2010 attributable to differences in coefficients is 66 versus 74 percent with versus without imputation; and the percentage of the gap in 2012 attributable to differences in coefficients is 73 percent with or without imputation.

In contrast, the measured discrimination in the DER data with versus without imputations varies across years. In 2006 and 2012, the percentage of the gap unexplained is larger by 2 to 3 percentage points with imputation versus without imputation. In 2009, the unexplained gaps are nearly the same with and without imputed data. In 2010 the percentage of the gap unexplained is 3 percentage points smaller with imputation versus without imputation.

Since it is not *always* the case that discriminatory measures are smaller with imputed data, we cannot conclude that there is a systemic bias in the use of imputation in the measurement of discrimination using CPS or DER data. The claim that imputation masks discrimination may appear relevant for some years in either CPS or DER data.

Conclusion

This paper presents data describing the black-white earnings gap before, during, and after the Great Recession. We examine changes in the black-white earnings gap over that period. Additionally, we examine variation in the gap by data source. Specifically, we are interested in understanding how self-reported earnings might influence the gap – either by masking true differences in black-white earnings or incorrectly explaining the gap through age, education, or sector of employment.

We find evidence that the CPS potentially masks the magnitude of difference in the earnings gap when the gap is calculated using a sample of CPS where earnings and full imputations of individuals are included. However, when we exclude imputed data, the earnings gap as measured with CPS self-reports and IRS administrative records more closely aligns. While we are still investigating the reasons for this, potential explanations could lie in the imputations themselves, in the fact that the sample changes with the more restrictive analysis of non-imputed individuals, or some other factor.

In the continuation of this work, we will conduct a more nuanced analysis of the imputations. For the purposes of research, our findings are relevant to the community because they highlight the fact that self-reported earnings data have the potential to produce flawed inequality estimates, perhaps masking the true level of discrimination and overestimating the true

influence of age, education, and employment sector on the earnings gap. We advise inequality researchers to proceed with caution when using this data and to consider using administrative earnings data for analysis.

References

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

Arias, Omar, Gustavo Yamada, and Luis Tejerina. 2004. "Education, Family Background, and Racial Earnings Inequality in Brazil." *International Journal of Manpower* 25:3/4, 355.

Bayer, Patrick, and Kerwin Kofi Charles. 2016. "Divergent Paths: Structural Change, Economic Rank, and the Evoluation of Black-White Earnings Differences, 1940-2014." *NBER Working Paper 22797*.

Bollinger, Christopher R., Barry T. Hirsch, Charles M. Hokayem, and James P. Ziliak. "Trouble in the Tails? What We Know about Earnings Nonresponse Thirty Years after Lillard, Smith, and Welch." *Unpublished Working Paper*, September 2015.

Cunningham, Wendy, and Joyce P. Jacobsen. 2008. "Earnings Inequality Within and Across Gender, Racial, and Ethnic Groups in Four Latin American Countries." *World Bank Policy Research Working Paper 4591*.

Darity, William A. Jr., Samuel L. Myers, Jr., and Chanjin Chung. 1998. "Racial Earnings Disparities and Family Structure." *Southern Economic Journal* 65:1, 20-41.

Gabriel, Paul E. 2004. "Differences in Earnings, Skills and Labour Market Experience Among Young Black and White Men." *Applied Economic Letters* 11, 337-341.

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

Hokayem, Charles, Trivellore Raghunathan, and Jonathan Rothbaum. “Sequential Regression Multivariate Imputation in the Current Population Survey Annual Social and Economic Supplement.” *SEHSD Working Paper 2016-XX*, May 2, 2016.

Masters, Stanley H. 1974. “The Effect of Educational Differences and Labor-Market Discrimination on the Relative Earnings of Black Males.” *The Journal of Human Resources* 9:3, 342-360.

Meyer, Bruce D., Wallace K. C. Mok, and James X. Sullivan, 2015. “Household Surveys in Crisis.” *Journal of Economic Perspectives* 29:4, 199-226.

O’Neill, June. 1990. “The Role of Human Capital in Earnings Differences Between Black and White Men.” *Journal of Economic Perspectives* 4:4, 25-45.

Posey, Kirby G. “Household Income: 2015.” *U.S. Census Bureau Technical Report ACSBR/15-02*, September 2016.

Proctor, Bernadette D., Jessica L. Semega, and Melissa A. Kollar. “Income and Poverty in the United States: 2015.” *U.S. Census Bureau Technical Report P60-256(RV)*, September 2016.

Smith, James P. and Finish R. Welch. 1977. “Black-White Male Wage Ratios: 1960-70.” *The American Economic Review* 67:3, 323-338.

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