Handling Respondent Rounding of Wages Using the IRS and CPS Matched Dataset

Minsun K. Riddles, Sharon L. Lohr, and J. Michael Brick, Westat, and
Patrick T. Langetieg, J. Mark Payne, and Alan H. Plumley, Internal Revenue Service

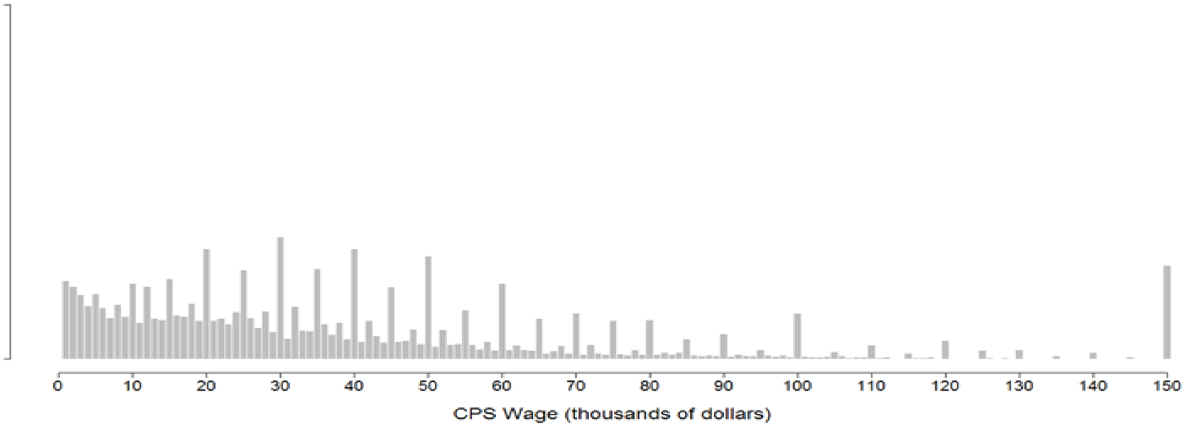
1. Introduction

Every year, the Internal Revenue Service (IRS) collects wage information from employers for services performed by employees. Employers are required to file a W-2 form for each employee from whom income tax, Social Security tax, or medicare tax was withheld; and for each employee who would have been subject to income tax withholding had he or she not avoided withholding by claiming additional allowances or exemption from withholding (Internal Revenue Service (2010)).

The Current Population Survey (CPS) also collects information on wage income. The CPS, a monthly household survey conducted by the U.S. Census Bureau for the U.S. Bureau of Labor Statistics, is the primary source of monthly labor force statistics and provides information on the economic and social well-being of the population of the United States. The target population is the civilian noninstitutionalized population 16 years of age and older. The CPS March Annual Social and Economic Supplement collects supplemental data on health insurance coverage, previous year's income from all sources, work experience, receipt of noncash benefits, and other topics. (U.S. Census Bureau (2012)).

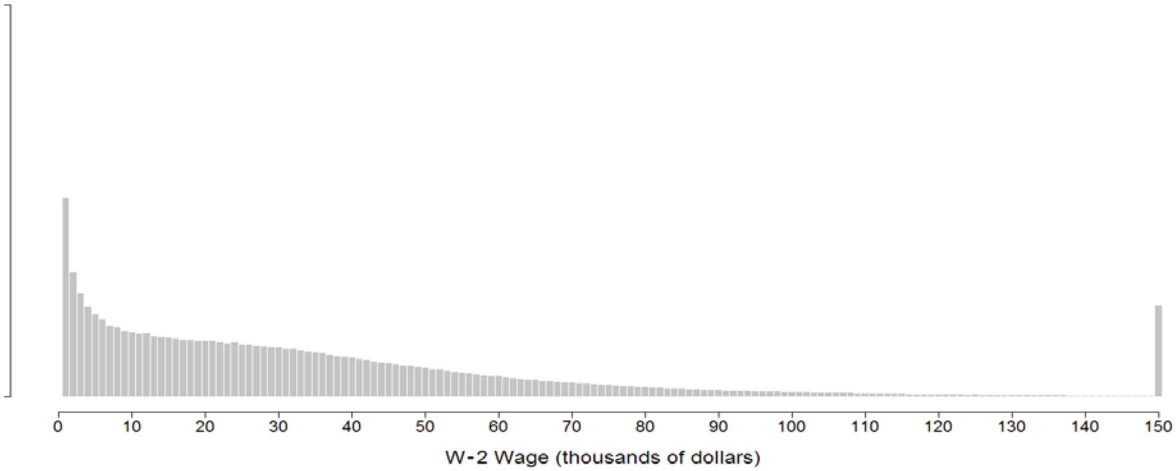
The CPS respondents are asked to report income from various sources. The CPS wage income is calculated from the responses to questions about earnings from an employer of their longest job and any other wage and salary earnings in the previous year. Figure 1 shows a weighted histogram of the 2011 CPS reported total wage income. The survey weights are used to calculate the estimated frequencies of reported total wage income in the histogram bins of [\$0.01–\$1,000], [\$1,000.01–\$2,000], [\$2,000.01–\$3,000], …, and [\$148,000.01–\$149,000], [\$149,000.01+). Persons with wages above \$150,000 are topcoded into the highest bin. The estimated frequencies exclude the CPS respondents with imputed wage income. They also exclude CPS respondents who reported no wage income (54 percent of the CPS respondents). Since the CPS wage income is self-reported by respondents, it is subject to measurement errors. In particular, some respondents may round the reported wage to a multiple of \$5,000, \$10,000, or some other "round number." The spikes in the histogram in Figure 1 represent a mixture of true values (some persons in the \$60,000 bin have actual wage income between \$59,000 and \$60,000) and rounded values (e.g., a person with wages of \$57,314 may round that value to \$60,000).

FIGURE 1. Weighted Histogram of CPS Reported Wages in the 2011 CPS March Supplement (Excluding \$0 Wages)



On the other hand, the histogram for the IRS W-2 wage income of Tax Year 2010 in Figure 2 does not have the same pattern of spikes because the wage income from a W-2 form is unrounded.

FIGURE 2. Histogram of W-2 Wages in the W-2 Population, Tax Year 2010



The populations for the CPS and the IRS data are not the same. The CPS is intended to produce statistics on the U.S. civilian noninstitutionalized population ages 16 and older living in housing units. The IRS W-2 data contain information for everyone with reportable wages, including persons under age 16, in institutions, or outside the United States. The CPS data contain many records for persons not represented in the IRS data because they have no reportable income. Some, but not all, of these persons report zero wages on the CPS.

However, it is possible to match records from the CPS to the IRS data using the unique Protected Identification Key (PIK) assigned to each person by the Census Bureau.[1] Table 1 shows the number of records in the 2011 CPS that were matched vs. not matched to the IRS data (rounded to the nearest 1,000), as well as the number of records with no wage amount reported. Table 1 suggests that the wage distribution of the non-matched cases is quite different from the wage distribution of the matched cases. Matching the two datasets provides an opportunity to check the validity of assumptions pertaining to the CPS data. For this paper, only the wages of the matched respondents were used for the analysis, and our attention is restricted to possible rounding of wages reported on the CPS. We do not consider other types of potential measurement errors.

We thus consider two data sets. The first consists of wage data for all W-2 forms for Tax Year 2010, which we call the W-2 population. No weights are used for estimates from this data set because it contains the entire population. The second consists of CPS respondents whose records can be linked to a W-2 form, which we call the CPS-IRS matched cases. For the second data set, the CPS weights are used to calculate histogram frequencies. There are two sources of wage information for the matched cases: the W-2 wage information, and the CPS self-reported wages.

TABLE 1. 2011 CPS (Tax Year 2010) Respondents by PIK and Matching Statuses

CPS respondents	Number of respondents	Percentage (unweighted)	Number of respondents reporting 0 wage	Percentage of respondents reporting 0 wage
Have a PIK	**164,000**	**92%**	**88,000**	**53%**
PIK matched to IRS	78,000	44%	8,000	10%
PIK not matched to IRS	86,000	48%	80,000	93%
No PIK	**15,000**	**8%**	**9,000**	**61%**
Total	**179,000**	**100%**	**97,000**	**54%**

[1] The Census Bureau assigns the PIKs to both Census and IRS data based on name, address, age, and other characteristics. See Jones and O'Hara (2014) and Wagner and Layne (2012).

Many researchers have studied problems of estimating the underlying distribution of unobserved unrounded values when the reported data have heaping in the context of: reporting the number of cigarettes smoked (Heitjan and Rubin (1991)), where some respondents may round to the nearest 20, which is the number of cigarettes in a pack; age (Heitjan and Rubin (1990)), where children's ages may be rounded to a multiple of 6 or 12 months; job-search duration (Torelli and Trivellato (1993)), in which respondents tend to report durations that are a multiple of 12 months; or medical measurements, which may be rounded to the nearest integer (Wright and Bray (2003)). Various methods have been developed for estimating the original distribution of values from rounded data (Riddles and Lohr (2015a,b); Zhang and Heitjan (2007); Drechsler, *et al.* (2015); Zinn and Wurbach (2015)).

The goals of this research are to: (1) find a density to fit the histogram of the W-2 wages from the W-2 population; (2) model the rounding mechanism for the self-reported wage income in the CPS and estimate a smoothed density for the CPS self-reported wages that accounts for the rounding; and (3) compare the distribution of CPS self-reported wages with the distribution of W-2 wages for the matched cases.

In Section 2, we explore the W-2 wage distribution in the W-2 population in order to find a model for the underlying distribution of unrounded wages. We present the distribution of the W-2 wages for the CPS-IRS matched cases in Section 3 and check if the model found in Section 2 also fits the W-2 wages for the CPS-IRS matched cases. Section 4 describes the model for how CPS respondents round and presents the results using the underlying distribution of unrounded wages and the model for the rounding mechanism. Section 5 summarizes the findings.

2. W-2 Wage in the W-2 Population

The W-2 wage income in the Tax Year 2010 W-2 population (N = 150,963,474) was explored to find a model capturing the distribution of W-2 wage income. Only 0.0003 percent of the W-2 population have zero W-2 wage income. W-2 wages of zero were excluded from this analysis. To preserve the confidentiality of records, the information on W-2 wages was summarized by IRS as the frequency distribution shown in Figure 2. The values were categorized in bins of width $1,000 up to $150,000 and bins of width $5,000 for wage income greater than $150,000 and topcoded at $300,000. For this analysis, high wages were topcoded at $150,000.

Let Y represent the maximum value of the bin for each value in the W-2 dataset. For example, each point in the first histogram bin, representing W-2 wages in [$0.01–$1,000], is given a Y value of $1,000. The variable Y has a discrete distribution, and we are interested in the smooth distribution that would fit the distribution of true wages (before binning), X. Assuming a parametric distribution for the true wage income (X) of $f(x \mid \boldsymbol{\theta})$, the density of Y can be written by integrating the density of X over each histogram bin, giving

$$g(y \mid \boldsymbol{\theta}) = \begin{cases} F(y \mid \boldsymbol{\theta}) - F(y-1{,}000 \mid \boldsymbol{\theta}) & \text{if } y \le 150{,}000 \\ 1 - F(y-1{,}000 \mid \boldsymbol{\theta}) & \text{if } y = 150{,}000, \end{cases}$$

where $y = \in \{\$1{,}000, \$2{,}000, \cdots, \$149{,}000, \$150{,}000\}$ and $F(\cdot \mid \theta)$ is the cumulative distribution function of $f(\cdot \mid \theta)$.

We assume that the true wages, X, are from a three component log-normal mixture distribution as follows:

$$f(x \mid \boldsymbol{\theta}) = \lambda_1 f_l(x \mid \mu_1, \sigma_1) + \lambda_2 f_l(x \mid \mu_2, \sigma_2) + (1 - \lambda_1 - \lambda_2) f_l(x \mid \mu_3, \sigma_3), \tag{1}$$

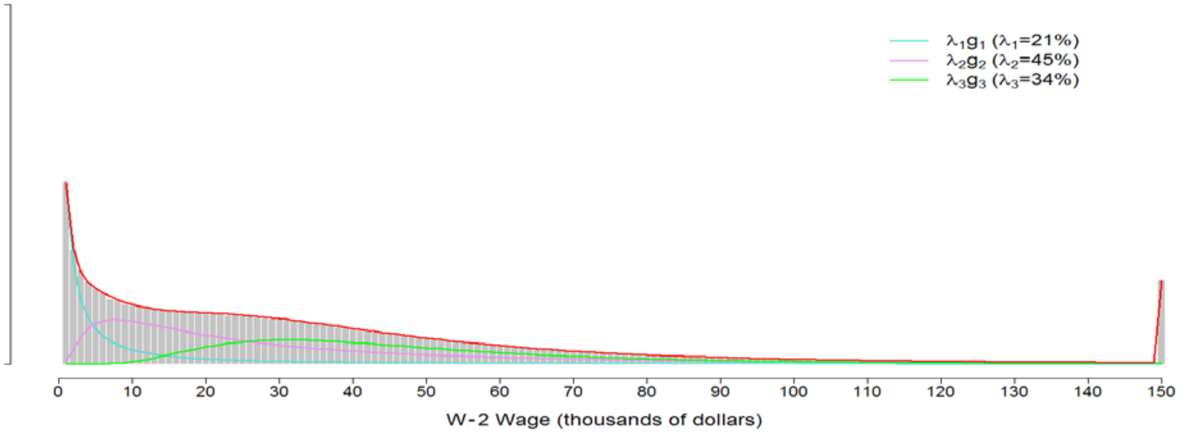
where $\boldsymbol{\theta} = (\lambda_1, \lambda_2, \mu_1, \sigma_1, \mu_2, \sigma_2, \mu_3, \sigma_3)$ and $f_l(\cdot \mid \mu, \sigma)$ is the probability density function of a lognormal distribution with parameters μ and σ. This model was selected after considering simpler models such as a single lognormal distribution and a two-component lognormal mixture distribution: neither of these was flexible enough to capture the density of W-2 wages between $10,000 and $30,000. Throughout this paper, $f(\cdot \mid \boldsymbol{\theta})$ in (1) is used as the distribution of the underlying "true" wage income.

Given the underlying distribution, $f(\cdot\,|\,\boldsymbol{\theta})$ in (1), the distribution of Y can be rewritten as:

$$g(y\,|\,\boldsymbol{\theta}) = \begin{cases} \sum_{j=1}^{3}\lambda_j \left\{F_l(y\,|\,\mu_j,\sigma_j) - F_l(y-1{,}000\,|\,\mu_j,\sigma_j)\right\} & \text{if } y \le 150{,}000 \\ \sum_{j=1}^{3}\lambda_j \left\{1 - F_l(y-1{,}000\,|\,\mu_j,\sigma_j)\right\} & \text{if } y = 150{,}000, \end{cases} \tag{2}$$

where $\lambda_3 = 1 - \lambda_1 - \lambda_2$ and $F_l(\cdot\,|\,\mu,\sigma)$ is the cumulative distribution function of a lognormal distribution with parameters μ and σ. The maximum likelihood estimates for $\boldsymbol{\theta}$ are found using computational methods described in Riddles and Lohr (2015a). The parameter estimates for all models are given in Table 2 in Section 5. Figure 3 presents the distribution of the W-2 wages in the Tax Year 2010 W-2 population with the fitted lognormal mixture distribution and its three estimated mixture components. This suggests that the distribution in (1), $f(\cdot\,|\,\boldsymbol{\theta})$ fits the W-2 wage population distribution very well.

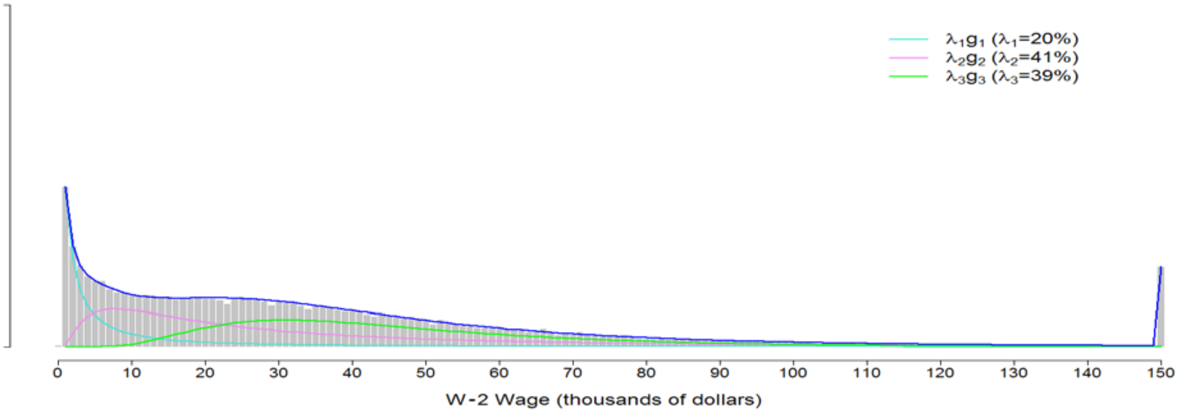
FIGURE 3. Histogram of W-2 Wages in W-2 Population, With Fitted Lognormal Mixture Distribution and Its Three Mixture Components, Tax Year 2010



3. W-2 Wages for the CPS-IRS Matched Cases

In this section, we use the model described in Section 2 to estimate the distribution of W-2 wages for the CPS-IRS matched cases (n = 78,199). Note that the CPS weights are incorporated in the analysis in this section, because the binned frequencies are calculated using the weights. Although the weighted distribution (Figure 4) is not as smooth as the distribution in the Tax Year 2010 W-2 population (Figure 3), the two distributions appear similar. The distribution $g(y\,|\,\boldsymbol{\theta})$ in (1) is fitted for this dataset by finding the maximum likelihood estimates of $\boldsymbol{\theta}$. Figure 4 presents the distribution of the W-2 wages for the matched cases with the fitted lognormal mixture distribution and its three mixture components. The parameter estimates differ slightly from those for the entire W-2 population, but Figure 4 shows that the general form of the three-component lognormal mixture model also fits the W-2 wage distribution for the CPS-IRS matched cases.

FIGURE 4. Weighted Histogram of W-2 Wages Among CPS-IRS Matched Cases, With Fitted Lognormal Mixture Distribution and Its Three Mixture Components, Tax Year 2010 (March 2011 CPS Supplement)



4. CPS Reported Wages for CPS-IRS Matched Cases

Not surprisingly, the distribution of CPS reported wages for CPS-IRS matched cases is not smooth and its weighted histogram (Figure 5) shows heaping at multiples of \$5,000, \$10,000, and \$50,000. Also, some heaping at \$12,000 and \$18,000 is present; perhaps this is from persons who round their monthly income to the nearest \$1,000 or \$1,500. In order to take into account heaping at multiples of \$5,000, \$6,000, \$10,000, and \$50,000 in the CPS reported wages, we specify a rounding mechanism as follows:

$$G = \begin{cases} 0 & \textit{if rounded to nearest } b_0 = 1{,}000 \\ 1 & \textit{if rounded to nearest } b_1 = 5{,}000 \\ 2 & \textit{if rounded to nearest } b_2 = 6{,}000 \\ 3 & \textit{if rounded to nearest } b_3 = 10{,}000 \\ 4 & \textit{if rounded to nearest } b \ = 50{,}000. \end{cases}$$

Using the methodology developed in Riddles and Lohr (2015a), we assume that heaping is caused only by rounding, and the rounding mechanism, G, depends on the true value, X, with a nonproportional odds cumulative logit model as follows:

$$\pi_0(x \mid \boldsymbol{\gamma}) = \frac{1}{1+\exp(\gamma_1+\gamma_2 x)},$$

$$\pi_1(x \mid \boldsymbol{\gamma}) = \frac{\exp(\gamma_1+\gamma_2 x)}{1+\exp(\gamma_1+\gamma_2 x)} - \frac{\exp(\gamma_3+\gamma_4 x)}{1+\exp(\gamma_3+\gamma_4 x)},$$

$$\pi_2(x \mid \boldsymbol{\gamma}) = \frac{\exp(\gamma_3+\gamma_4 x)}{1+\exp(\gamma_3+\gamma_4 x)} - \frac{\exp(\gamma_5+\gamma_6 x)}{1+\exp(\gamma_5+\gamma_6 x)},$$

$$\pi_3(x \mid \boldsymbol{\gamma}) = \frac{\exp(\gamma_5+\gamma_6 x)}{1+\exp(\gamma_5+\gamma_6 x)} - \frac{\exp(\gamma_7+\gamma_8 x)}{1+\exp(\gamma_7+\gamma_8 x)},$$

and

$$\pi_4(x \mid \boldsymbol{\gamma}) = \frac{\exp(\gamma_7+\gamma_8 x)}{1+\exp(\gamma_7+\gamma_8 x)},$$

where $\pi_g(x \mid \boldsymbol{\gamma}) = P(G = g \mid x, \boldsymbol{\gamma})$.

The goal of this section is to estimate the underlying (unrounded) distribution of CPS-reported wages, when only the reported rounded values are available in the data. We assume that the underlying distribution of the true values, X, can be fit by a mixture of three lognormal density functions of the form in equation (1). We also assume that the value reported by a respondent, Z, is obtained because the respondent uses one of the possible rounding mechanisms on the true value, X. This results in a density for the reported values, Z, that depends on the parameters $\boldsymbol{\theta}$ from the assumed density of the true values, $f(x|\boldsymbol{\theta})$, and also depends on parameters used to estimate the rounding mechanism, $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \ldots, \gamma_8)$, as follows:

$$h(z|\boldsymbol{\gamma},\boldsymbol{\theta}) = \begin{cases} \sum_{g=0}^{4} \int_{0}^{z+b_g/2} \pi_g(t|\boldsymbol{\gamma}) f(t|\boldsymbol{\theta}) dt & \text{if } z = 0 \\ \sum_{g=0}^{4} \mathrm{I}_g(z) \int_{z-b_g/2}^{z+b_g/2} \pi_g(t|\boldsymbol{\gamma}) f(t|\boldsymbol{\theta}) dt & \text{if } 0 < z \leq 150{,}000, \\ \sum_{g=0}^{4} \int_{z-b_g/2}^{\infty} \pi_g(t|\boldsymbol{\gamma}) f(t|\boldsymbol{\theta}) dt & \text{if } z = 150{,}000, \end{cases}$$

where $f(\cdot|\boldsymbol{\theta})$ is defined in (1) and $\mathrm{I}_g(z)$ is equal to 1 if z is a multiple of b_g and 0 otherwise.

Figure 5 shows the fitted distribution for $h(\cdot|\gamma,\boldsymbol{\theta})$using the maximum likelihood estimate of $(\gamma, \boldsymbol{\theta})$. This is the estimated distribution of the rounded, self-reported values. Figure 5 suggests that the assumed model $h(\cdot|\gamma,\boldsymbol{\theta})$ fits the distribution of the CPS-reported wages well and captures heaping at multiples of \$5,000, \$6,000, \$10,000, and \$50,000.

Figure 6 shows the estimated density of the unrounded values for the CPS wages, superimposed on the histogram. The assumed rounding mechanism smooths out the spikes in the histogram, allowing comparison of the estimated density with the density fit to the W-2 wages.

FIGURE 5. Weighted Histogram of CPS Reported Wages Among CPS-IRS Matched Cases, With Fitted Distribution, Tax Year 2010 (March 2011 CPS Supplement)

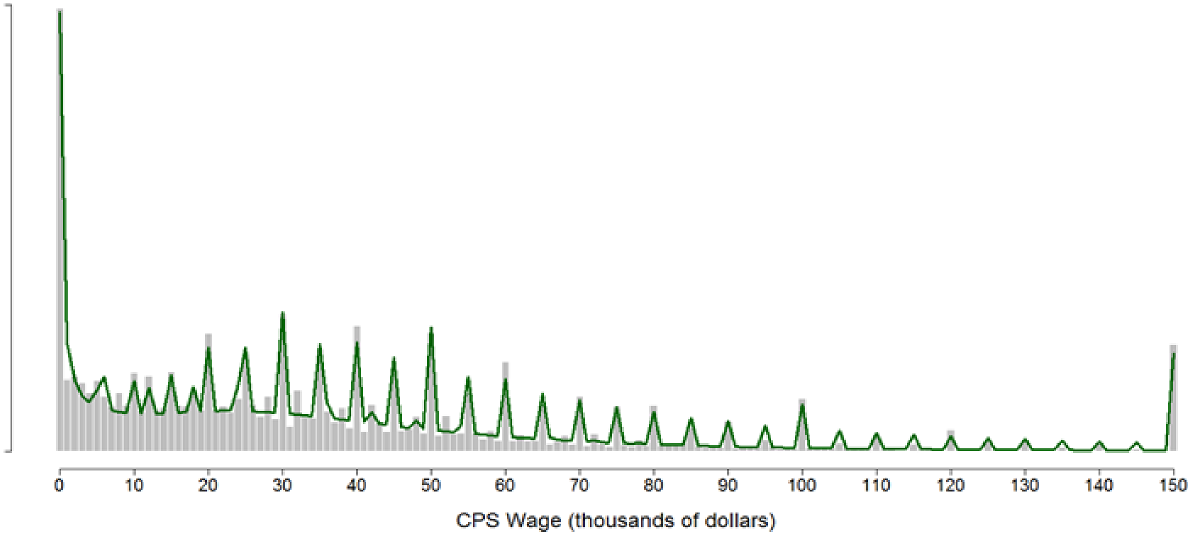
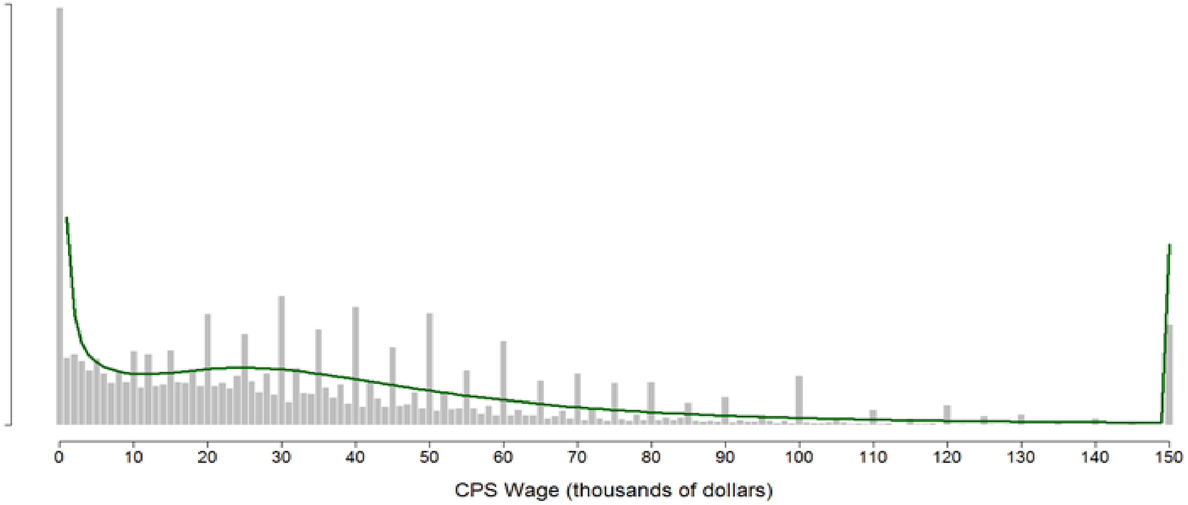


FIGURE 6. Weighted Histogram of CPS Reported Wages Among CPS-IRS Matched Cases, With Estimated Density of the Unrounded CPS Wages, Tax Year 2010 (March 2011 CPS Supplement)



5. Summary

Table 2 presents the three sets of parameter estimates for $\boldsymbol{\theta}$ in the underlying distribution, $f(x|\boldsymbol{\theta})$. The parameter estimates are very similar but not identical across the three sets (from W-2 wages in the W-2 population, W-2 wages for the CPS-IRS matched cases, and CPS reported wages for the matched cases). All of these estimated distributions have three lognormal components, as in equation (1). Note that a lognormal distribution has mean $\exp(\mu+\sigma^2/2)$, median $\exp(\mu)$, and mode $\exp(\mu-\sigma^2)$, and this allows us to identify the components of the mixture distributions on the graphs. For the W-2 wage population data, the first component has $\mu = 8.1$ and $\sigma = 1.762$ and corresponds to the blue line in Figure 3, accounting for approximately 21 percent of the total density. The second component, with $\mu = 10.043$ and $\sigma = 1.073$, corresponds to the pink line in Figure 3 and accounts for approximately 45 percent of the total density. The third component, with $\mu = 10.619$ and $\sigma = 0.54$, corresponds to the green line in Figure 3 and accounts for the remaining approximately 34 percent of the density. These are empirical mixtures, designed to fit the empirical data distribution, and they do not correspond to specific subpopulations.

TABLE 2. Parameter Estimates for θ Using W-2 Wages in the W-2 Population, W-2 Wages for CPS-IRS Matched Cases, and CPS Reported Wages for CPS-IRS Matched Cases, Tax Year 2010 (March 2011 CPS Supplement)

	Wage:	W-2 wage	W-2 wage	CPS wage
	Source:	W-2 population	CPS-IRS matched	CPS-IRS matched
Parameters	λ_1	0.210	0.197	0.197
	λ_2	0.452	0.409	0.409
	λ_3	0.338	0.394	0.394
	μ_1	8.100	8.140	8.144
	σ_1	1.762	1.792	1.928
	μ_2	10.043	10.506	10.407
	σ_2	1.073	1.086	1.121
	μ_3	10.619	10.616	10.616
	σ_3	0.540	0.542	0.544

The distribution for W-2 wages was estimated using the three sets of parameter estimates in Table 2: (1) using the W-2 wages in the Tax Year 2010 W-2 population from Section 2; (2) using the W-2 wages for the matched cases from Section 3; and (3) using the CPS reported wages for the matched cases from Section 4. Figure 7 shows the distribution of W-2 wage income in the W-2 population with these three estimated distributions. The comparisons of densities show some differences between the estimates of W-2 wages from the W-2 data and the self-reported wage data from the CPS.

To see the effect of these differences on estimated percentiles of wages, we look at the cumulative distribution function (CDF) for each estimate. Figure 8 presents the three estimated CDFs of unrounded wage income: (1) using the W-2 wages in the Tax Year 2010 W-2 population; (2) using the W-2 wages for the matched cases; and (3) using the CPS reported wages for the matched cases. The five horizontal lines in light gray correspond to values of 0.1, 0.25, 0.5, 0.75, and 0.9, respectively. The wage value where the horizontal line meets the estimated CDF is the estimate of the corresponding percentile. For example, the wage values where the horizontal line for 0.5 meets the three estimated CDFs are the three estimated medians.

Figures 7 and 8 consistently show that the two estimated distributions based on W-2 wages are very close to each other, and the estimated distribution based on CPS wages fits the W-2 wages for the matched cases fairly well, but underestimates the density of wages under $12,000 and overestimates the density of wages of $150,000 or more. These differences result in differences for the estimated percentiles of the wage distribution. For example, Figure 8 shows that the estimated median wage from the self-reported CPS data is approximately $4,000 higher than the estimated median wage for the same persons with the W-2 data.

In this paper, we fit a smooth density to the histogram of the W-2 wages from the W-2 population. We then adopted the form of that density for the “true” (before rounding) values in the CPS wage data, and estimated the parameters for that density along with the parameters for the rounding mechanism for the self-reported wage income in the CPS. We have applied a full likelihood-based approach developed in Riddles and Lohr (2015a) to estimate the distribution of unrounded wages using the CPS reported wages with both the model for W-2 wages and the model for the rounding mechanism.

The model smooths out the distribution of the CPS self-reported wages, and allows comparison with the estimated density from the W-2 data. Although the estimated distribution of the CPS self-reported wages is close to that of the W-2 data overall, there are differences that indicate there may be measurement errors that have not been captured by the models. Additional research is needed to investigate sources of differences between the smoothed CPS estimates and the W-2 wage estimates.

FIGURE 7. Histogram of W-2 Wages in W-2 Population, With Estimated Distributions: (1) Using W-2 Wages in W-2 Population (Red); (2) Using W-2 Wages for Matched Cases (Blue); and (3) Using CPS Reported Wages for Matched Cases (Green), Tax Year 2010 (March 2011 CPS Supplement)

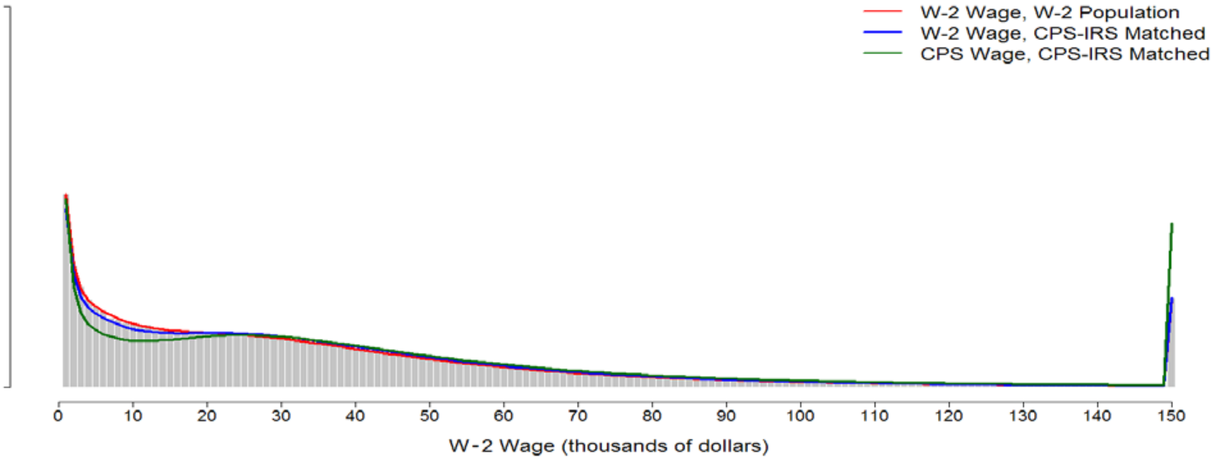
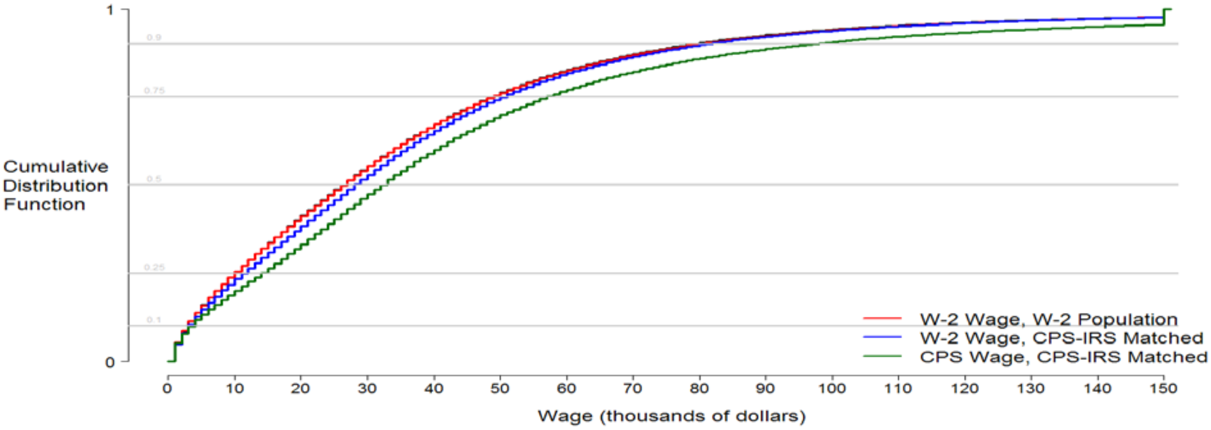


FIGURE 8. Estimated Cumulative Distribution Functions of Unrounded Wages: (1) Using W-2 Wages in W-2 Population (Red); (2) Using W-2 Wages for Matched Cases (Blue); and (3) Using CPS Reported Wages for Matched Cases (Green), Tax Year 2010 (March 2011 CPS Supplement)



References

Drechsler, J., Kiesl, H., and Speidel, M. (2015), “MI Double Feature: Multiple Imputation To Address Nonresponse and Rounding Errors in Income Questions,” *Austrian Journal of Statistics*, 44, 59–71.

Heitjan, D. F., and Rubin, D. B. (1990), “Inference from Coarse Data via Multiple Imputation with Application to Age Heaping,” *Journal of the American Statistical Association*, 85, 304–314.

— (1991), “Ignorability and Coarse Data,” *Annals of Statistics*, 19, 2244–2253.

Internal Revenue Service (2010), “2010 Instructions for Forms W-2 and W-3,” available at https://www.irs.gov/pub/irs-prior/iw2w3--2010.pdf.

Jones, M. R., and O’Hara, A. (2014), “Do Doubled-Up Families Minimize Household-Level Tax Burden?,” *2014 IRS Research Bulletin*, Publication 1500, 181–203.

Riddles, M. K. and Lohr, S. L. (2015a), “A Maximum Likelihood Approach to Estimating Parametric Distributions with Rounded Responses,” Technical Report submitted to Internal Revenue Service.

— (2015b), “A Mixture Model Approach for Heaped Data with Rounded Responses and True Spikes,” Technical Report submitted to Internal Revenue Service.

Torelli, N., and Trivellato, U. (1993), “Modelling Inaccuracies in Job-search Duration Data,” *Journal of Econometrics*, 59, 187–211.

U.S. Census Bureau (2012), “March 2011: Annual Social and Economic (ASEC) Supplement Technical Documentation,” available at https://www.census.gov/prod/techdoc/cps/cpsmar11.pdf.

Wagner, D., and Layne, M. (2012), “Person Identification Validation System (PVS): Applying the Center for Administrative Records Research and Applications’ Record Linkage Software,” Washington, DC: Center for Administrative Records Research and Applications Internal Document, U.S. Census Bureau.

Wright, D., and Bray, I. (2003), “A Mixture Model for Rounded Data,” *Journal of the Royal Statistical Society Series D*, 52, 3–13.

Zhang, J. and Heitjan, D. F. (2007), “Impact of Nonignorable Coarsening on Bayesian Inference,” *Biostatistics*, 8, 722–743.

Zinn, S., and Würbach, A. (2015), “A Statistical Approach To Address the Problem of Heaping in Self-Reported Income Data,” *Journal of Applied Statistics* [online], DOI: 10.1080/02664763.2015.1077372, available at http://www.tandfonline.com/loi/cjas20.