Total Error and Variability Measures with Integrated Disclosure Limitation for Quarterly Workforce Indicators and LEHD Origin Destination Employment Statistics in OnTheMap

Kevin L. McKinney Andrew S. Green Lars Vilhuber
John M. Abowd*

December 16, 2017

*Kevin McKinney (kevin.l.mckinney@census.gov) is Senior Economist, U.S. Census Bureau. Andrew Green (asg248@cornell.edu) is Economist, U.S. Census Bureau, and Economics Ph.D. student at Cornell University, Lars Vilhuber (lars.vilhuber@cornell.edu) is Senior Research Associate and Executive Director of the Labor Dynamics Institute at Cornell University and Economist (IPA), U.S. Census Bureau. John Abowd (john.maron.abowd@census.gov) is Associate Director for Research and Methodology and Chief Scientist, U.S. Census Bureau, Edmund Ezra Day Professor of Economics, Statistics and Information Science, Cornell University (ILR School), and Director, Labor Dynamics Institute, Cornell University ILR School.

Abstract

We report results from the first comprehensive total quality evaluation of five major indicators in the U.S. Census Bureau's Longitudinal Employer-Household Dynamics (LEHD) Program Quarterly Workforce Indicators (QWI): total employment, beginning-of-quarter employment, full-quarter employment, total payroll, and average monthly earnings of full-quarter employees. Beginning-of-quarter employment is also the main tabulation variable in the LEHD Origin-Destination Employment Statistics (LODES) workplace reports as displayed in OnTheMap (OTM). The evaluation is conducted by generating multiple threads of the edit and imputation models used in the LEHD Infrastructure File System. These threads conform to the Rubin (1987) multiple imputation model, with each thread or implicate being the output of formal probability models that address coverage, edit, and imputation errors. Design-based sampling variability and finite population corrections are also included in the evaluation. We derive special formulas for the Rubin total variability and its components that are consistent with the disclosure avoidance system used for QWI and LODES/OTM workplace reports. These formulas allow us to publish the complete set of detailed total quality measures for QWI and LODES. The analysis reveals that the five publication variables under study are estimated very accurately for tabulations involving at least 10 jobs. Tabulations involving three to nine jobs have quality in the range generally deemed acceptable. Tabulations involving zero, one or two jobs, which are generally suppressed in the QWI and synthesized in LODES, have substantial total variability but their publication in LODES allows the formation of larger custom aggregations, which will in general have the accuracy estimated for tabulations in the QWI based on a similar number of workers.

Keywords: Multiple imputation; Total quality measures; Employment statistics; Earnings statistics; Total survey error.

Acknowledgements

Portions of Appendix A are based on an unpublished technical memo dated February 1, 2011 by John Abowd, Henry Hyatt, Mark Kutzbach, Erika McEntarfer, Kevin McKinney, Michael Strain, Lars Vilhuber, and Chen Zhao. Liliana Sousa was an important contributor to the work documented in the Appendix. We received helpful comments from Erika McEntarfer and John Eltinge. Sara Sullivan edited the final manuscript. Abowd and Vilhuber acknowledge direct support from the U.S. Census Bureau (prior to Abowd's appointment) and NSF Grants SES-0922005, BCS 0941226, TC-1012593, and SES-1131848. This research uses data from the U.S. Census Bureau's Longitudinal Employer-Household Dynamics Program, which was partially supported by the following National Science Foundation Grants: SES-9978093, SES-0339191 and ITR-0427889; National Institute on Aging Grant AG018854; and grants from the Alfred P. Sloan Foundation. Any opinions and conclusions expressed herein are those of the authors and do not necessarily represent the views of the U.S. Census Bureau. All results have been reviewed to ensure that no confidential information is disclosed. The data produced by the analysis described in this paper have been released for public use and may be found at `http://doi.org/10.3886/E100590V1`.

1 Introduction and Summary

We compute the first comprehensive estimates of total error and variability for two Longitudinal Employer-Household Dynamics (LEHD) products from the U.S. Census Bureau: the Quarterly Workforce Indicators (QWI), which are public-use tables displayed in QWI Explorer, and the workplace-based LEHD Origin-Destination Employment Statistics (LODES), which are the public-use tables displayed in OnTheMap (OTM) when a workplace report is requested. These labor market indicators are produced from a comprehensive integrated administrative record system known as the LEHD Infrastructure File System, which is based primarily on the linkage between employers and employees provided by state-regulated unemployment insurance (UI) wage records. The theoretical universe to which these records correspond is all statutory jobs in the economy – private and public (excluding federal employees).[1] There is also a benchmark census of all such jobs in the universe: the Quarterly Census of Employment and Wages (QCEW) from the Bureau of Labor Statistic (BLS). We use this census, which is also integrated into the LEHD Infrastructure File System as the finite population that the QWI and LODES tabulations estimate. In principle, the published indicators are subject to errors from coverage, sampling, edit, and imputation. By addressing all of these sources of error in our assessment of total variability, we have created the first comprehensive total quality measures for these data.

Coverage errors are addressed in two ways. First, each wage record is linked to the associated employer record from the putative universe of employers (QCEW). When there is a link, estimated employment from the two sources is compared. A tentative weight is constructed to adjust the LEHD Infrastructure File System estimate of employment. When there is not a link, an entity is added to the LEHD infrastructure version of the QCEW, called the Employer Characteristics File (ECF), to account for this absence. At the end of the processing, a final weight is computed that benchmarks all employment to the BLS published state-level employment totals for the same universe. The effect of this proce-

[1] At the time this evaluation was undertaken, federal employees were not covered in QWI and LODES.

dure is to transmit the coverage errors into the edit and imputation procedures used to complete the firm level tabulation variables when there is a linkage failure in the data integration. Details of these record-linkage procedures are discussed in Abowd and Vilhuber (2005), Benedetto et al. (2007), and Abowd et al. (2009).

Every job in the universe must have completed data for all the publication variables. The LEHD Infrastructure File System has a fully-integrated collection of probability models that generate multiply-imputed values for all missing data items in the system. Most details are supplied in Abowd et al. (2009) – in particular, the models for imputing missing demographic and workplace characteristics.[2] The system uses the methods first proposed by Rubin (1987) and expanded in Little and Rubin (2002) for analyses using multiply-imputed missing data. The total variance statistics described in this paper are based on specially adapted versions of the Rubin measures generated using the approved QWI disclosure avoidance method: input noise-infusion as described in Abowd et al. (2009) and Abowd et al. (2012).

Users of these total variability measures have several options. The measures are intended to provide the information needed to construct approximate confidence intervals at all levels of stratification for five key publication statistics: total employment, beginning-quarter employment, full-quarter employment, total payroll, and average monthly earnings of full-quarter employees. We give detailed guidance on how to use our results to calculate confidence intervals for arbitrary published employment totals and earnings.

The Rubin measures are also designed to summarize the extent to which the variability due to the edit and imputation procedures, as distinct from the variability due to sampling in the underlying data, contributes to total variation. Total variability consists of both between-implicate variance generated primarily by edit and imputation, and within-implicate variance, which consists primarily of variability due to sampling. However, the sampling variance is small since in principle we should have the population of firms and

[2]Abowd et al. (2009) does not document the replacement to the demographic variable imputation methods that were incorporated in 2010. Those methods are documented in Appendix A.

jobs.

We also distinguish and account for variability due to sampling and structural zeros. In the language of Bishop et al. (1975), a structural zero occurs whenever there is no reported activity in a cell – that is, no business exists in the cell – and a sampling zero occurs when the cell is at risk to have positive employment (because a business exists) but does not. We treat the probability that a job will be classified in a particular detailed category of the publication tables as potentially random within a fixed population of state jobs. This set of assumptions allows us to model the equivalent of sampling variability as if it were generated by a particular multinomial model.

All five indicators we study are published every quarter in the QWI, stratified by ownership, sub-state geography, detailed industry, worker age, gender, race, ethnicity, and education. The publication tables also cross-classify many of these stratifiers. Beginning-of-quarter employment is the primary tabulation variable in LODES for display in OnTheMap, which is released annually with many of the same stratifiers as in the QWI and sub-state geography down to the block level. Constructing measures of total variability for these indicators is complicated by three related factors. First, the QWI and LODES are produced in separate production streams although they share the core LEHD Infrastructure File System and, therefore, are subject to the same sources of variation. Neither production stream saves all the inputs required to calculate total variability. Second, the QWI are revised quarterly, and revised indicators are released for the complete history of the series. Third, the workplace-based statistics produced by the LEHD program for both QWI and LODES/OTM use a confidentiality protection system based on input noise-infusion that constrains the calculation of total variability measures and complicates the release of these measures in a user-friendly format.

Because the QWI production system does not store the implicate threads needed to compute the total variability statistics, the analysis in this paper is based on a re-creation of the production statistics from the research files corresponding to a particular vintage of

the QWI. The research code does not exactly match the production code. In particular, there are discrepancies in the counts between the production and research values of the statistics for which we compute total variability measures. Even if the research code did exactly reproduce the publication statistics from one release of QWI, the next quarter's release would not agree exactly for most of the historical data because of the continuous-revision design of QWI. The user must take care when calculating confidence intervals for the published values using the total variability measures tabulated here. There are two available strategies, both of which are discussed in this paper. The user can download a table of total variability measures with the same structure as the tabulations for which confidence intervals are required. In this case, there will be some discordance between the value of the indicator found in the publication tables and the value that was used to calculate the total variability measures. We document when these discrepancies can be important: unsurprisingly, mostly for cells with small tabulation counts. We also provide detailed tables that can be used directly to construct approximate confidence intervals.

Overall, these comprehensive measures of the total quality of QWI and LODES tabulations for five critical variables provide substantial evidence that the system is producing reliable data. This summary discusses the qualitative results for the main employment indicator used by both QWI and LODES/OTM, beginning-of-quarter employment.

Both QWI and LODES/OTM were designed to allow detailed sub-state geography and industry tabulations. Such a system, of necessity, must be robust to the presence of many cells with very small tabulations and many zeros. We document that the vast majority of zeros result from no reported activity, meaning that the value is exactly zero and is treated as a structural zero. Since QWI and LODES are population tabulations, structural zeros have no variability, which is imposed in our analysis. Some zeros are estimated, and those zeros have total variability. Cells with small published employment totals (for any of the employment measures) do have substantial estimated total variability.

The smallest tabulation values (cells containing counts of one or two) often have

90% confidence intervals of less than plus or minus one, so that they sometimes include zero and three. These values are usually suppressed in QWI but they are released in LODES/OTM. The suppression in QWI is justified because the full hierarchical tabulation is published, reducing the need for custom aggregations; however, QWI users and QWI Explorer, the Census Bureau's own analysis tool for these data, do generate custom tabulations. These custom tabulations must populate the cells with suppressed items using some algorithm. There are no suppressions in LODES/OTM, which completes the data using a synthetic data model based on the posterior predictive distribution of small cell counts (one or two) within a given tract stratified by most of the variables for which LODES tabulations are published. Regardless of the model used, there is still substantial uncertainty in these small tabulations, as our results confirm. Almost all 90% confidence intervals are tighter than the interval zero to five, while the vast majority are less than plus or minus two. Publication of these small tabulations in spite of their substantial relative uncertainty is justified by the flexibility they allow for generating custom tabulation areas, most of which end up with much larger estimated employment totals. These custom tabulations would be substantially biased by using zero as the estimate when the publication value of a component is suppressed.

For cells where the tabulations are in the range of three to nine, our results indicate that the 90% confidence interval is rarely wider than plus or minus three, and for most tables is less than plus or minus one. For cells where the tabulations are in the range of 10 or more, it makes more sense to summarize the results in terms of percentage variation; i.e., use the coefficient of variation implied by the total variability measure and the estimated count. For tabulations in the range of 10 to 99 jobs, the 90% confidence intervals are rarely larger than plus or minus 25% and are usually in the range of plus or minus 10% to 25%. For tabulations in the range 100 to 999, the widest 90% confidence intervals are plus or minus 20%, and the vast majority of cells in this range have confidence intervals of plus or minus less than 10%. For the largest tabulation areas, 1,000 or more, the

widest 90% confidence intervals are approximately plus or minus 5%, and the intervals are usually in the range of plus or minus 1.5%.

The other dimension along which we assess the total variability is the Rubin missingness ratio, which quantifies the proportion of the total variability that arises from the multiple imputation procedures. This is also known as "fraction of missing information" as in Little and Rubin (2002). The complement of the Rubin missingness ratio measures the proportion of total variability that it is due to sampling and other intrinsic sources of randomness in the indicator; that is, the proportion of total variability that would remain if no records required any edits or imputation. As we noted above, the edit and imputation procedures used in QWI and LODES/OTM also induce variability due to sub-state coverage errors.

The Rubin missingness ratio provides a reasonable way to assess the effects of data edits and imputations for both demographic characteristics (age, gender, race, ethnicity, and education) and workplace characteristics (industry and county). When age and gender are the only two stratifiers used in the publication table, missing data account for about 44% of total variability. When race and ethnicity are the only two stratifiers, missing data account for between 80% and 95% of total variability. When gender and education are the only stratifiers in the publication tables, missing data account for over 95% of total variability. When workplace industry and county are the only stratifiers in the publication tables, missing data account for between 0% and 80% of total variability. It is important to note that even when the Rubin missingness ratio is large, the 90% confidence intervals implied by the total variability measure remain as summarized above. The missingness ratios are a guide to where improvements in the data quality either through the use of measured data from other sources or through better imputation algorithms can reduce total variability the most.

In addition to contributing to the literature on variance estimation using multiply imputed data, this study contributes to the growing body of literature on total survey er-

ror. Biemer (2010) defines total survey error as the "accumulation of all errors that may arise in the design, collection, processing, and analysis of survey data." The total error estimates undertaken in this study address errors due to coverage, sampling, edit, and imputation. This accounts for almost all sources of error due to the representational properties of the survey (Groves et al., 2004; Groves and Lyberg, 2010). This study contributes to a recent, if more mature, literature which uses administrative data to evaluate existing surveys, as well as an emerging literature which seeks to asses the total quality of administrative data itself.[3] The final assessment adheres closely to the best practices enumerated across many statistical agencies when applied to current data products.[4]

The remainder of this paper is organized as follows. Section 2 provides background on the missing data problem and the methods for multiply imputing worker and establishment characteristics. Section 3 provides formal models for estimating total variability and its associated components in a manner that is fully consistent with the required disclosure avoidance procedures. To the best of our knowledge, these formulas have never been derived or published before. Section 4 discusses the detailed results and provides guidance for computing confidence intervals. Section 5 concludes.

2 Background on QWI, LODES, and the Multiply Imputed Characteristics

The QWI are a public-use data product of the U.S. Census Bureau. Every quarter, local labor market statistics are released by worker demographics, workplace geography, and other employer characteristics. Unlike many labor force statistics derived from surveys of workers or employers, the QWI are produced from job-based administrative data, where a job is the link of a statutory employee to a statutory employer. This linkage allows the

[3]See Mulry and Keller (2017) and Reid et al. (2017) for two such examples.

[4]See Eurostat (2014) and Horrigan et al. (2014) for examples of total quality frameworks applied to other data outputs.

QWI to provide tabulations of labor force statistics by worker and employer characteristics, such as county employment by firm size and gender. In addition, the unique identifiers for the employer and worker allow the QWI to tabulate longitudinal statistics, such as hires, separations, and turnover.

The LODES are similar to the QWI in that they originate from the same job-based frame. However, the LODES data provide geographic detail for both place of work and place of residence, but only release a subset of the labor force statistics in the QWI, and are published annually with statistics derived using the first day of the second quarter of the year (April 1st) as the reference date. The core employment variable, beginning-of-quarter employment, called B below, is used for both the QWI and LODES tabulations.[5]

The QWI and LODES are based on the LEHD Infrastructure File System. The original production version of this system is documented in Abowd et al. (2009). The LEHD infrastructure files are made possible through the Local Employment Dynamics state-federal partnership where participating states provide the U.S. Census Bureau quarterly extracts of earnings records from their respective UI systems as well as an extract from the QCEW, as specified by a similar federal-state cooperative arrangement between the states and the BLS.

The UI earning records are used to construct a job-based frame for the QWI and LODES. An in-scope job occurs when a worker produces at least one dollar of UI-covered earnings at a non-federal establishment in a given quarter. The LEHD Infrastructure File System then combines this information with additional survey and administrative data to derive individual characteristics such as age, gender, place of birth, race, ethnicity, and education, as well as establishment characteristics, such as workplace address and North American Industrial Classification System (NAICS) codes. The LEHD Infrastructure File System was developed using model-based edit and imputation procedures. Every missing data element has been multiply-imputed using an integrated set of models described

[5]Publication tables for the QWI can be found here: `http://qwiexplorer.ces.census.gov/`. Publication maps for LODES/OTM can be found here: `http://onthemap.ces.census.gov/`.

in Abowd et al. (2009). There are 10 implicates for every missing item. Implicates are denoted by $l = 1, \ldots, L$. The missing data models for most of the variables used in this paper, including birth date, gender, race, ethnicity, education, workplace geography, workplace NAICS, firm age, and firm size, have been substantially improved and modified since the 2009 article was written. Because the LEHD Infrastructure File System is rebuilt every quarter from all historical records, the analysis in this paper incorporates all of those model improvements.

The LEHD program receives unemployment insurance records from states without any individual or workplace characteristics. They provide the basis for constructing a comprehensive frame of jobs. Individual characteristics are added to the job frame from a variety of Census Bureau surveys and federal administrative data. The five worker characteristics are birth date, sex, race, ethnicity, and education, each of which is part of an integrated multiple imputation model. This model is based on discrete categories for each variable. The imputation process starts with variable(s) having the least missing data, taking advantage of what is commonly known as a monotone missing data pattern, although in this case it is approximate. At each stage of the modeling, imputations from the earlier stages are used as conditioning information for the active variable. Missing birth date and sex are imputed in the first stage. In the second stage, missing race and ethnicity are imputed. Finally, missing education is imputed. Appendix A contains detailed documentation of the individual characteristics imputation.

In addition to worker characteristics, a separate process imputes the workplace characteristics for each record in the job frame. Workplace characteristics are based on associating an establishment with each job spell in the LEHD data. A job spell is the collection of quarterly unemployment insurance records that pertain to the same worker and employing firm with an interruption of no more than four quarters. States deliver the unemployment insurance wage records each quarter, which form the core of the job frame, at the employee-firm-state level, where a "firm" is defined as a state unemployment insurance ac-

count number. In addition, the states provide a quarterly list of all known establishments owned by the firm within a state as part of the QCEW extract. This list includes establishment characteristics such as industry and geography, as well as the employment counts in the reference week for each establishment for each month in the quarter. However, with the exception of Minnesota, explicit identifiers directly linking an employee to an establishment do not exist. In order to produce labor market statistics for detailed industries and geographies, the link associating a worker with an establishment is multiply imputed.[6]

The QWI and the workplace component of LODES are confidentiality protected using an input noise-infusion method applied to the underlying micro-data. Every establishment (identifiers: SEIN, SEINUNIT) in the database has been assigned a unique noise factor, δ_j, where j indexes establishments that satisfy the conditions documented in Abowd et al. (2009, 2012). We refer to this unique input noise factor as the "fuzz factor" for the establishment and employer. The method for applying this fuzz factor to the publication statistics depends upon whether the publication statistic is based on a magnitude (including employment counts for an establishment), ratio, or other more complicated statistic. In addition, small magnitude values in the QWI are suppressed with the flag "5: Does not meet Census Bureau publication standards" and significantly distorted publication values are labeled with the flag "9: Significantly distorted." In LODES, values that would be suppressed in QWI are synthesized using a probability model that is based on the posterior predictive distribution of the suppressed values conditional on tract-level establishment employment data.

The total variability statistics described in this paper apply to data for all private employers and the current all-employer category in the QWI and LODES data, which ex-

[6]The data from Minnesota are used to fit a hierarchical Bayesian model of establishment assignment. The probability of an employee working at a given establishment is estimated in this hierarchical structure with the first part conditioning on the employment sizes of all establishments at the firm (SEINUNITs within SEIN), and the second part conditioning on the distance between an employee's residence and each establishment. The model is fit jointly on each of three firm size categories. The estimated model parameters and the size distribution of establishments within the firm are used to generate 10 draws of feasible establishments for each job. For further details see Abowd et al. (2009).

cludes federal employees. Statistics that include only federal employees are covered by a different protection procedure. Statistics that aggregate all-employer data (excluding federal employment) with federal employment data must combine the two types of data from their respective public-use releases.

We extend the QWI noise-infusion methods to cover the protection of the Rubin total variance measure for statistics based upon multiply-imputed missing data. This measure combines the conventional quality measure for published statistics – the design-based sampling variance, corrected for *ex post* departures from design and finite populations – and a measure that captures the contribution of the model-based missing data imputation procedures: the between-implicate variance of the publication statistic.

3 Noise-Infusion Protected Total Variance Measures

This section derives the formulas for noise-infusion protected Rubin total variance measures. To the best of our knowledge, these formulas have never been derived or published before. We restrict our analysis to five core labor force statistics published in the QWI:

- Beginning-of-quarter employment, B, which is equal to the sum of all workers who had positive earnings at an establishment in the current quarter as well as the previous quarter.
- Full-quarter employment, F, which is defined as the sum of all workers who had positive earnings at an establishment in the current quarter in addition to the previous and subsequent quarters.
- Average monthly earnings of full-quarter employees, Z_W3.
- Total flow-employment, M, defined as the sum of all workers who have positive earnings at an establishment at any time in the quarter.
- Total payroll, $W1$, which is the total earnings earned by workers in a quarter.

Beginning-of-quarter employment for quarter two (April 1-June 30) is also the primary tabulation variable in LODES/OTM.

The relevant population is a state.[7] At the state level, the QCEW measure of all employment (excluding federal workers) is considered the population. Quarterly weights for the QWI benchmark B to the QCEW month-1 employed population. All statistics defined below are calculated for a given state-year-quarter. Similar to the actual QWI, total variability statistics are produced for the period beginning in 1990, quarter one (1990:1). The total variability measures discussed in this paper refer to the QWI release labeled R2012Q4, which covers 1990:1 through 2012:1. All states except Massachusetts, North Carolina, and Colorado are included in the R2012Q4 release.[8]

We adopt, without modification, the noise-infusion methodology described in Abowd et al. (2009) and elaborated in Abowd et al. (2012), to which the reader is referred for more details. The system adds multiplicative noise to tabular output produced from the LEHD Infrastructure File System. The multiplicative noise factors for each establishment are drawn from a two-sided symmetric ramp distribution centered at the value one. The draws from the distribution distort the original input by at least a minimum percentage, and by no more than a maximum percentage. Both of these values are Census confidential. This system is a substantial generalization of the method originally developed by Evans et al. (1998). As applied in the production of the QWI and LODES/OTM, the release statistics are dynamically consistent – the same noise factor is used for an establishment in every quarter of data.

The system also provides protection to employers as well as establishments – all establishments for the same employer within a given state have noise distortion factors on the

[7]For simplicity, we include Washington, D.C. when we say "state."

[8]The schema for the QWI at the R2012Q4 release are described at `https://lehd.ces.census.gov/data/schema/v3.5/`. The schema for later QWI releases changed, at the time of writing, the latest schema documentation was available at `https://lehd.ces.census.gov/data/schema/V4.1.3/lehd_public_use_schema.html`. Availability for each state varies both historically and at any point in time, see `https://lehd.ces.census.gov/doc/QWI_data_notices.pdf` (archival version) for available data for each state. The estimated total variability measures described in this paper can be downloaded here: `http://doi.org/10.3886/E100590V1`.

same side of unity. The system can provide protection to magnitude measures, the only problem considered by Evans et al. (1998), ratios, and differences. Employment counts within demographic categories are treated as magnitudes. The protection method for ratios requires that the publication tables include either two magnitudes (e.g., total employment and total payroll) or one magnitude and one ratio (e.g., total employment and average quarterly earnings). We use the ratio form of the QWI noise-distortion protection below.[9]

Multiplicative noise infusion provides confidentiality protection in the following sense. The originally reported values of the tabulation variables are never used in the formation of the magnitudes (establishment-level counts and sums) and ratios that are tabulated. The input noise infusion insures that for every micro-data record tabulated, there is a strictly positive percentage difference between the value used in tabulation and the true confidential value. Tabulations based upon a small number of establishments (at the limit one) or a small number of employees (at the limit one) contain uncertainty induced by the distribution of the noise factor. This uncertainty limits a user's ability to infer attributes to within a range that is confidential. Finally, the physical location of a workplace is not treated as confidential because it is defined as the location where an employee must report for work, and is therefore public. While the protection system is not formally private in the sense of Dwork et al. (2006), it does satisfy the necessary conditions in Dinur and Nissim (2003) for resistance to database reconstruction attacks. See Haney et al. (2017) for a formal privacy analysis of this protection mechanism.

3.1 Population Definitions

To calculate the components of total variance, every quarter we require estimates of the total population, N_{WB}, and the total sample size, N_{UB}. To be consistent with the

[9]We do not use the protection method for differences in this paper.

overall data protection scheme, we must calculate these from the fuzzed data as

$$N_{WB} = \sum_{\forall j} B_j^U w_j \delta_j \equiv \sum_{\forall j} B_j^* \quad \text{and} \tag{1}$$

$$N_{UB} = \sum_{\forall j} B_j^U \delta_j \equiv \sum_{\forall j} B_j^{U*} \,, \tag{2}$$

where B_j^U is the unweighted establishment-level beginning-of-quarter employment for establishment j, w_j is the QWI establishment weight, δ_j is the unique QWI establishment fuzz factor, B_j^* is the fuzzed-weighted establishment-level count, and B_j^{U*} is the fuzzed-unweighted count establishment-level count. Summing over all firms gives us estimates of N_{WB} and N_{UB} (excluding federal establishments). N_{WF}, N_{UF}, F_j^*, F_j^U, and F_j^{U*} are defined similarly for full-quarter employment, as well as N_{WM}, N_{UM}, M_j^*, M_j^U, and M_j^{U*} for total employment. The population estimate N_{WB} has been benchmarked to the QCEW month-1 employed population via the QWI weights. This procedure is also discussed in Abowd et al. (2009). There is no QCEW population count for full-quarter employment nor total employment. However, N_{WF} and N_{WM} are treated here as the appropriate estimate of the population total for F and M, respectively. Since Z_W3 is computed over the same set of input records as F, its fuzzed-weighted and fuzzed-unweighted population and total sample counts are identical to N_{WF} and N_{UF}. $W1$ is calculated using earnings for all workers, thus, N_{WM} and N_{UM} are the correct population and sample size for this statistic.

In principle, for all the missing data models, there should not be any between-implicate variance in N_{WB}, N_{UB}, N_{WF}, N_{UF}, N_{WM}, and N_{UM} because missing records are corrected using the weights and only missing items on actual records are imputed. Therefore, it should not make any difference which implicate is used to compute these population and sample totals. We computed population totals separately for each implicate and attempted to verify the absence of between-implicate variation in the total fuzzed-weighted and fuzzed-unweighted counts. In practice, there is a small amount of between-implicate variance in the population totals – less than 0.04% for B and less than 0.03% for

F as measured by the coefficient of variation. The results are tabulated by state in Appendix Table A.11 for beginning-of-quarter population, and in Appendix Table A.12 for the full-quarter population. The between-variance measures are also computed for each establishment type (private and all, excluding federal). These results are also displayed in Appendix Tables A.11 and A.12. Between-implicate variation in the sub-population totals is consistent with the benchmarking but is also minimal.

3.2 Total Variability Models for B, F, and M

Let B_k be any cross-classification of beginning-of-quarter employment such that $N_{WB} = \sum_{\forall k} B_k$. For each implicate l, the fuzzed-weighted count for category k is computed as

$$B_k^{(l)*} = \sum_{(i,j)\in\{\text{def } k\}} b_{i,j}^{(l)} w_j \delta_j \tag{3}$$

where $b_{i,j}^{(l)}$ is the LEHD infrastructure indicator variable that defines person i as a beginning-of-quarter employee of establishment j in the l^{th} implicate (implicitly, for date t), $\{\text{def } k\}$ is the set that defines membership in category k for the pair (i, j), and w_j is the QWI weight for establishment j. $F_k^{(l)*}$ and $M_k^{(l)*}$ are defined comparably using the LEHD infrastructure indicator variables $f_{i,j}^{(l)}$ and $m_{i,j}^{(l)}$, respectively, and the same weight and fuzz factor as in the equation for $B_k^{(l)*}$.

For each implicate, the estimated proportion of N_{WB} represented by $B_k^{(l)*}$ in each cell k is

$$p_k^{(l)*} = \frac{B_k^{(l)*}}{N_{WB}} \, . \tag{4}$$

The estimated count in cell k can be rewritten as

$$B_k^{(l)*} \equiv c_k^{(l)*} = N_{WB} \times p_k^{(l)*} \, . \tag{5}$$

The released statistics are the averages taken over the implicates

$$B_k^* \equiv \bar{c}_k* = \frac{1}{L}\sum_{l=l}^{L} c_k^{(l)*} \quad \text{and} \tag{6}$$

$$\frac{B_k^*}{N_{WB}} \equiv \bar{p}_k* = \frac{1}{L}\sum_{l=l}^{L} p_k^{(l)*} \ . \tag{7}$$

Exactly comparable formulas are used for F_k^* and M_k^*.

For each implicate, the finite-population-corrected, *ex-post*-design-weighted sampling variance of the proportion is estimated by assuming that the counts are sampled from a multinomial population and that the missing infrastructure records (equivalent of non-response or coverage errors) are corrected by the QWI weights. Only fuzzed inputs are used in the calculation. Hence, the estimator for the within-implicate variance of the proportion is

$$vp_k^{(l)*} = \left(\frac{p_k^{(l)*}\left(1-p_k^{(l)*}\right)}{N_{UB}}\right)\left(\frac{N_{WB}-N_{UB}}{N_{WB}-1}\right) \ . \tag{8}$$

For each implicate, the finite-population-corrected, *ex-post*-design-weighted sampling variance of the count is estimated with

$$vc_k^{(l)*} = N_{WB}^2\left(\frac{p_k^{(l)*}\left(1-p_k^{(l)*}\right)}{N_{UB}}\right)\left(\frac{N_{WB}-N_{UB}}{N_{WB}-1}\right) \ . \tag{9}$$

Again, only fuzzed inputs are used.[10]

Notice that the finite population correction (the last term) is not at the cell level. Population counts are only known for beginning-of-quarter employment. Due to problems with population counts in small cells when the relevant population is not beginning-of-quarter employment, we use the state level population correction for all cells. This implicitly assumes that the ratio of the sample to the population is the same as beginning-of-quarter

[10]In the next production of these total variability estimates, we will apply the correction for clustering workers in establishments within firms as described in Cochran (1977, pp. 64-68).

employment, where the population is known.

The Rubin between-variances for the proportions and counts are

$$bp_k^* = \frac{1}{L-1}\sum_{l=1}^{L}\left(p_k^{(l)*} - \bar{p}_k^*\right)^2 \quad \text{and} \tag{10}$$

$$bc_k^* = \frac{1}{L-1}\sum_{l=1}^{L}\left(c_k^{(l)*} - \bar{c}_k^*\right)^2 \ . \tag{11}$$

The Rubin average within-variances for the proportions and counts are

$$\bar{vp}_k^* = \frac{1}{L}\sum_{l=1}^{L} vp_k^{(l)*} \quad \text{and} \tag{12}$$

$$\bar{vc}_k^* = \frac{1}{L}\sum_{l=1}^{L} vc_k^{(l)*} \ . \tag{13}$$

The Rubin total variances are

$$tvp_k^* = \bar{vp}_k^* + \left(\frac{L+1}{L}\right) bp_k^* \quad \text{and} \tag{14}$$

$$tvc_k^* = \bar{vc}_k^* + \left(\frac{L+1}{L}\right) bc_k^* \ . \tag{15}$$

For completeness, we also calculate the Rubin missingness ratio as

$$mrp_k^* = \frac{\left(\frac{L+1}{L}\right) bp_k^*}{tvp_k^*} \ , \tag{16}$$

and similarly for mrc_k^*.

All formulas for full-quarter employment and total employment, F and M, are comparable – substituting $f_{i,j}^{(l)}$ for $b_{i,j}^{(l)}$, $F_k^{(l)}$ for $B_k^{(l)}$, N_{WF} for N_{WB}, and N_{UF} for N_{UB} in the case of F, with analogous substitutions for M. Because N_{WF} and N_{WM} are not benchmarked by the QCEW but are based on the weights for beginning of-of-quarter employment, there may be negative finite population corrections that we replaced with the smallest positive

finite-population correction factor based on B.[11]

3.3 Total Variability Model for Z_W3

The cells for $Z\_W3_k^*$ are the same mutually-exclusive and exhaustive cells as used for F_k^*. For any implicate l, the fuzzed-weighted estimate of average monthly earnings is calculated as

$$Z\_W3_k^{(l)*} = \frac{1}{F_k^{(l)}} \sum_{(i,j)\in\{\text{def } k\}} z\_w3_{i,j}^{(l)} w_j \delta_j, \tag{17}$$

where $F_k^{(l)}$ is the unfuzzed-weighted full-quarter employment for cell k. To compute the sampling variance of $Z\_W3_k^{(l)*}$, we use the fuzzed-weighted uncorrected sum of squares, calculated as

$$uss_k^{(l)*} = \sum_{(i,j)\in\{\text{def } k\}} \left(z\_w3_{i,j}^{(l)}\right)^2 w_j \delta_j \ . \tag{18}$$

For each implicate, the finite-population-corrected, *ex-post*-design-weighted sampling variance of the average monthly earnings for full-quarter employed workers is estimated with

$$vz_k^{(l)*} = \frac{1}{F_k^{(l)u}} \left(\frac{uss_k^{(l)*}}{F_k^{(l)}} - \left(Z\_W3_k^{(l)*}\right)^2 \right) \left(\frac{N_{WF} - N_{UF}}{N_{WF} - 1} \right) \tag{19}$$

where $F_k^{(l)u}$ is the unfuzzed-unweighted count of full-quarter employment in cell k, and $vz_k^{(l)*}$ is only computed when $F_k^{(l)}$ is positive. Notice that the formula for the within-variance for each implicate is a conditional sampling variance, given membership in cell k. In all cases unfuzzed-weighted values are used in the denominator and fuzzed values (weighted or unweighted) are used in the numerator. This is consistent with the approved QWI noise-infusion system and prevents cancellation of the fuzz-factor when only one establishment populates the cell. Because the average, $Z\_W3_k^{(l)*}$, is computed according to equation 17 and the mean uncorrected sum of squares is computed using the same denom-

[11] This procedure is essentially the same as the method used for finite population corrections in the American Community Survey (Starsinic, 2011).

inator as $Z\_W3_k^{(l)*}$, the term $\left(\frac{uss_k^{(l)*}}{F_k^{(l)}} - \left(Z\_W3_k^{(l)*}\right)^2\right)$ in equation 19 can be negative. This situation arises for small values, generally less than three, of $F_k^{(l)}$ when the discrepancy between the fuzzed count $F_k^{(l)*}$ and the unfuzzed count $F_k^{(l)}$ is relatively large. When this happens, the term $\left(\frac{uss_k^{(l)*}}{F_k^{(l)}} - \left(Z\_W3_k^{(l)*}\right)^2\right)$ is set to zero attributing all variation to the between-implicate variance.

The quantities for the Rubin total variance can now be computed for $Z\_W3_k^{(l)*}$. The publication statistic is

$$Z\_W3_k^* \equiv z\_\bar{w}3_k^* = \frac{1}{L}\sum_{l=1}^{L} Z\_W3_k^{(l)*} \ . \tag{20}$$

The between-implicate variance is

$$bz_k^* = \frac{1}{L-1}\sum_{l=1}^{L}\left(Z\_W3_k^{(l)*} - z\_\bar{w}3_k^*\right)^2 \ . \tag{21}$$

The average within-implicate variance is

$$v\bar{z}_k^* = \frac{1}{L}\sum_{l=1}^{L} vz_k^{(l)*} \ . \tag{22}$$

Finally, the Rubin total variance is

$$tvz_k^* = v\bar{z}_k^* + \frac{L+1}{L} bz_k^* \tag{23}$$

We also calculate the Rubin missingness ratio for average monthly earnings of full-quarter employees using the formula equivalent to equation 16.

3.4 Total Variability Model for $W1$

The cells for total payroll, $W1_k^*$, are the same mutually-exclusive and exhaustive cells as used for M_k^*. For any implicate l, the fuzzed-weighted estimate of total payroll is

$$W1_k^{(l)*} = \sum_{(i,j)\in\{\text{def } k\}} w1_{i,j}^{(l)} w_j \delta_j \tag{24}$$

where $w1_{i,j}^{(l)}$ is the gross payroll in cell k. To compute the sampling variance of $W1_k^{(l)*}$, we use the average payroll per worker multiplied by an estimate of the number of workers in cell k, $W1_k^{(l)*} = M_k^{(l)*} \times Z\_W1_k^{(l)*}$. First, we require the fuzzed-weighted estimate of average quarterly earnings, which is calculated as

$$Z\_W1_k^{(l)*} = \frac{1}{M_k^{(l)}} \sum_{(i,j)\in\{\text{def } k\}} w1_{i,j}^{(l)} w_j \delta_j \tag{25}$$

where $M_k^{(l)}$ is the unfuzzed-weighted employment flow for cell k. We also have the fuzzed-weighted uncorrected sum of squares, calculated as

$$mss_k^{(l)*} = \sum_{(i,j)\in\{\text{def } k\}} \left(w1_{i,j}^{(l)}\right)^2 w_j \delta_j \tag{26}$$

For each implicate, the finite-population-corrected, *ex-post*-design-weighted sampling variance of total payroll is estimated with

$$vw_k^{(l)*} = \frac{\left(M_k^{(l)*}\right)^2}{M_k^{(l)u}} \left(\frac{mss_k^{(l)*}}{M_k^{(l)}} - \left(Z\_W1_k^{(l)*}\right)^2\right) \left(\frac{N_{WM} - N_{UM}}{N_{WM} - 1}\right) \tag{27}$$

where N_{WM} and N_{UM} are the fuzzed-weighted count and the fuzzed-unweighted counts of population flows, respectively. The denominator in the first term, $M_k^{(l)u}$, is the unfuzzed-unweighted cell count. The numerator of the first term scales the sample mean to give the sample variance of a count. Just as with $Z\_W3_k^*$, the middle term in 27 may be negative,

which we then set to zero and attribute all variance to between-implicate variance.

The quantities for the Rubin total variance can now be computed for $W1_k^*$. The publication statistic is

$$W1_k^* \equiv \bar{W1}_k^* = \frac{1}{L}\sum_{l=1}^{L} W1_k^{(l)*} \ . \tag{28}$$

The between-implicate variance is

$$bw_k^* = \frac{1}{L-1}\sum_{l=1}^{L}\left(W1_k^{(l)*} - \bar{W1}_k^*\right)^2 \ . \tag{29}$$

The average within-implicate variance is

$$\bar{vw}_k^* = \frac{1}{L}\sum_{l=1}^{L} vw_k^{(l)*} \ . \tag{30}$$

Just as in equation 23, the Rubin total variance is

$$tvw_k^* = \bar{vw}_k^* + \frac{L+1}{L} bw_k^* \ . \tag{31}$$

We also calculate the Rubin missingness ratio for average monthly earnings of full-quarter employees using the formula equivalent to equation 16.

3.5 Reconciling Total Variability Measures Using Published Values of B, F, M, Z_W3, and $W1$

Once we compute the five QWI statistics, we perform quality checks and modify the within- and between-variance so they are consistent with public-use values. For reasons previously discussed, we compute the final total variability statistics using a research process distinct from the production process used to compute the QWI public-use statistics.[12] The resulting QWI statistics differ in some circumstances from the official public-

[12]To recap, research computing uses a snapshot of a single collection of vintages of the LEHD infrastructure file system that were used to compute one release of the data, in this case R2012Q4. Some pro-

use statistics, with the most discord occurring in the smallest public-use cells. To scale the internally calculated total variability statistics to the publicly released statistics, we assume the coefficient of variation is equal in both the public-use and internally calculated total variability statistics. In order ensure the reasonableness of this assumption, we edit the coefficient of variation of the QWI statistic when it deviates substantially from similar cells within the same aggregation level, and with the same size QWI statistic.

For each table, we merge a public-use table of QWI statistics with our corresponding internal calculations of the five QWI statistics and their associated total variability measures. Next, we bin each internally calculated employment measure, respectively, by aggregation level and into percentiles of employment. We calculate the 5th and 95th percentiles of the coefficient of variation for each bin. Within each bin, we consider cells below the 5th percentile and above the 95th percentile of the coefficient of variation outliers, and we replace their within- and between-variance with the within- and between-variance of the median of coefficient of variation of the bin. We also replace the internal statistic with the value of the corresponding median of the coefficient of variation of the bin. Note that the public-use statistic is always preserved and is the reference statistic for all total variability measures. Appendix B provides a more detailed summary of the procedure.

Before computing the released total variability measures consistent with the public-use QWI, we account for, and flag, the presence of sampling zeros. The public-use QWI contains only cells where at least one statistic is computable for the given cell, which means there is at least one UI-covered job in that cell. The frame for the QWI, however, is establishments whether they have positive UI-covered jobs in a quarter or not. Thus, it is possible that a given cell will have no released QWI statistics, but nonetheless be at risk for positive employment. This is a sampling zero. In contrast, some cells will never have positive employment or observed firm activity, and we denote these structural zeros because they are not at risk to have any employment in the cell. We flag these two types of

duction system edits are not captured in this snapshot. Similarly, some research system edits are not reflected in the production system.

cells for advanced users and estimate variance measures for the sampling zeros. Appendix C gives a detailed summary of the procedure.

After checks for the quality of the final statistics, we create the released statistics using the edited data and their corresponding statistics when necessary. We only release total variability statistics for unsuppressed statistics in the public-use QWI data. When the public-use value is close to the internal research value that we calculate, we scale the within- and between-variance by the square of the ratio of the public-use statistic to the internally computed statistic. Otherwise we scale using a representative value from another bin, invoking the assumption of equal coefficients of variation within a cell. The total variance, missingness ratio, and degrees of freedom are recalculated from the scaled within- and between-variance. The final file contains the same identifiers, QWI statistics, and status flags as the public-use tables. In addition, it includes the five total variability statistics rounded to three significant digits whenever the public-use statistic is present. The only additional records in the total variability files beyond those in the public-use QWI correspond to the sampling zeros, for which we report variability measures as described in Appendix C. The original, unscaled total variability statistics are used whenever either the public-use or internally calculated statistic is zero.

4 Results

We summarize the results in Table 1 for all total employment, $EmpTotal$, Table 2 for all beginning-of-quarter employment, Emp, in Table 3 for all full-quarter employment, $EmpS$, Table 4 for all total payroll, $Payroll$, in Table 5 for all average monthly earnings of full-quarter employees, $EarnS$. Tables showing the same statistics for only private establishments are shown in Appendix Tables A.6 to A.10. In addition to summaries of the statistics defined above, we also summarize the distribution of the coefficient of total variation, which is the square root of the total variance divided by the estimated statistic for

$EmpTotal$, Emp, $EmpS$, $EarnS$, and $Payroll$. For Emp this formula is

$$cvc_k^* = \frac{\sqrt{tvc_k^*}}{Emp_k^*} \tag{32}$$

The same equation holds for the four other statistics using their respective total variances in the numerator and the corresponding statistic in the denominator.

Table 1: Summary of Total Variability of All Total Employment (*EmpTotal*) by Table and Count

Table and *EmpTotal* count range	Proportion of Cells	Number of Cells	Median Count	Median Total Variation	Median Rubin Missingness Rate (Percent)	Quantiles of Coefficient of Variation			Median Approximate 90% Confidence Intervals Margin of Error		
						5th	Median	95th	Median df	Count	Percent
All (Private plus State and Local)											
Age x Gender											
+1000	1.0000	46,480	91,515	8690.00	43.10%	0.0003	0.0010	0.0032	48	121	0.13%
Race x Ethnicity											
10-99	0.0181	632	56	51.55	96.70%	0.0837	0.1403	0.2568	9	10	19.40%
100-999	0.1223	4,263	443	415.00	95.60%	0.0265	0.0474	0.0932	9	28	6.56%
+1000	0.8596	29,965	14,956	6310.00	87.30%	0.0002	0.0041	0.0269	11	108	0.56%
Gender x Education											
+1000	1.0000	23,240	187,994	222000.00	96.80%	0.0012	0.0028	0.0079	9	652	0.39%
Industry x County											
zero measured value, after rounding	0.0026	8,225	0	0.31	94.90%	(a)	(a)	(a)	10	1	(a)
1-2	0.0000	134	1	0.39	80.10%	0.1664	0.3652	0.9434	14	1	49.21%
3-9	0.0132	41,946	7	0.51	0.00%	0.0593	0.1075	0.3871	9999	1	13.78%
10-99	0.2333	743,122	47	5.52	43.50%	0.0237	0.0537	0.1643	47	3	6.98%
100-999	0.4539	1,445,825	307	57.90	70.80%	0.0101	0.0238	0.0593	18	10	3.17%
+1000	0.2971	946,236	2,989	774.00	77.10%	0.0023	0.0080	0.0199	15	37	1.08%
Age x Gender x Industry x County											
zero measured value, after rounding	0.1672	7,973,123	0	0.21	95.20%	(a)	(a)	(a)	9	1	(a)
1-2	0.0049	234,460	2	0.38	72.70%	0.1527	0.3252	0.7382	17	1	43.36%
3-9	0.2165	10,324,414	5	0.85	66.20%	0.0864	0.1754	0.3944	20	1	23.24%
10-99	0.4014	19,140,564	27	5.33	71.40%	0.0367	0.0806	0.1803	17	3	10.75%
100-999	0.1737	8,279,653	224	52.30	75.60%	0.0137	0.0294	0.0610	15	10	3.95%
+1000	0.0363	1,728,489	1,982	482.00	76.00%	0.0041	0.0101	0.0197	15	29	1.35%
Race x Ethnicity x Industry x County											
zero measured value, after rounding	0.5635	19,553,448	0	0.20	95.50%	(a)	(a)	(a)	9	1	(a)
1-2	0.0062	216,005	2	0.70	92.10%	0.2688	0.6245	0.9354	10	1	85.69%
3-9	0.1400	4,856,222	5	2.34	89.50%	0.1334	0.3159	0.5944	11	2	43.07%
10-99	0.1653	5,735,093	26	10.10	86.20%	0.0431	0.1169	0.2692	12	4	15.85%
100-999	0.0886	3,073,969	254	75.20	80.50%	0.0132	0.0317	0.0736	13	12	4.29%
+1000	0.0364	1,262,815	2,573	745.00	79.60%	0.0031	0.0093	0.0210	14	37	1.25%
Gender x Education x Industry x County											
zero measured value, after rounding	0.0737	1,787,333	0	0.26	95.10%	(a)	(a)	(a)	9	1	(a)
1-2	0.0044	106,593	2	1.38	93.40%	0.4290	0.6538	0.9513	10	2	89.72%
3-9	0.1901	4,610,815	5	4.05	93.20%	0.2386	0.3783	0.6101	10	3	51.91%
10-99	0.4433	10,755,591	29	22.50	93.20%	0.0853	0.1597	0.2946	10	7	21.91%
100-999	0.2305	5,593,317	234	187.00	93.40%	0.0291	0.0566	0.0963	10	19	7.76%
+1000	0.0580	1,407,901	2,090	1770.00	93.70%	0.0084	0.0192	0.0318	10	58	2.63%

Notes: Total employment is defined as all jobs held by a worker at the same establishment during the quarter. Statistics are computed across all state-year-quarters within a table. The "All" category of establishments includes private as well as state and local government but excludes federal employment. All tables include all valid QWI age groups with the exception of any table including education, in which case only jobs with workers age 25 and older are included. For statistic definitions for total employment, please see their respective equations in the accompanying text: Count 6, Total Variation 15, Missingness Ratio 16, Coefficient of Variation 32. (a) Undefined value.

Table 2: Summary of Total Variability of All Beginning-of-Quarter Employment (*Emp*) by Table and Count

Table and *Emp* count range	Proportion of Cells	Number of Cells	Median Count	Median Total Variation	Median Rubin Missingness Rate (Percent)	Quantiles of Coefficient of Variation			Median Approximate 90% Confidence Intervals Margin of Error		
						5th	Median	95th	Median df	Count	Percent
All (Private plus State and Local)											
Age x Gender											
+1000	1.0000	45,712	70,233	5300.00	37.00%	0.0003	0.0010	0.0032	65	94	0.13%
Race x Ethnicity											
10-99	0.0258	883	51	39.50	96.50%	0.0793	0.1277	0.2664	9	9	17.66%
100-999	0.1489	5,105	454	326.00	95.10%	0.0198	0.0430	0.0830	9	25	5.95%
+1000	0.8253	28,296	12,858	4340.00	84.60%	0.0001	0.0038	0.0237	12	89	0.52%
Gender x Education											
+1000	1.0000	22,856	161,812	162000.00	96.80%	0.0012	0.0028	0.0079	9	557	0.39%
Industry x County											
zero measured value, after rounding	0.0056	17,598	0	0.28	95.50%	(a)	(a)	(a)	9	1	(a)
1-2	0.0001	257	2	0.44	78.30%	0.1443	0.3592	0.8972	14	1	48.31%
3-9	0.0203	63,664	7	0.43	0.00%	0.0546	0.1022	0.3814	9999	1	13.09%
10-99	0.2633	827,121	45	5.06	50.30%	0.0223	0.0529	0.1643	35	3	6.91%
100-999	0.4464	1,402,205	295	55.70	74.70%	0.0099	0.0240	0.0590	16	10	3.20%
+1000	0.2643	830,357	2,875	711.00	79.30%	0.0023	0.0080	0.0197	14	36	1.07%
Age x Gender x Industry x County											
zero measured value, after rounding	0.2011	9,317,087	0	0.20	95.80%	(a)	(a)	(a)	9	1	(a)
1-2	0.0051	234,090	2	0.36	74.20%	0.1371	0.3156	0.7273	16	1	42.19%
3-9	0.2246	10,406,647	5	0.77	67.90%	0.0794	0.1675	0.3873	19	1	22.24%
10-99	0.3842	17,797,008	27	4.99	74.40%	0.0351	0.0791	0.1793	16	3	10.58%
100-999	0.1547	7,165,326	222	48.90	77.90%	0.0131	0.0288	0.0603	14	9	3.87%
+1000	0.0303	1,405,442	1,945	437.00	77.60%	0.0039	0.0097	0.0192	14	28	1.31%
Race x Ethnicity x Industry x County											
zero measured value, after rounding	0.6023	20,718,981	0	0.19	96.00%	(a)	(a)	(a)	9	1	(a)
1-2	0.0056	191,678	2	0.67	92.70%	0.2632	0.6050	0.9028	10	1	83.01%
3-9	0.1288	4,431,864	5	2.16	90.10%	0.1256	0.3044	0.5799	11	2	41.50%
10-99	0.1514	5,208,590	26	9.27	86.80%	0.0402	0.1115	0.2610	11	4	15.20%
100-999	0.0805	2,767,906	251	69.40	82.00%	0.0126	0.0307	0.0710	13	11	4.15%
+1000	0.0314	1,081,496	2,506	673.00	81.30%	0.0030	0.0091	0.0204	13	35	1.23%
Gender x Education x Industry x County											
zero measured value, after rounding	0.0870	2,055,422	0	0.26	95.70%	(a)	(a)	(a)	9	1	(a)
1-2	0.0049	114,711	2	1.37	94.20%	0.4269	0.6496	0.9421	10	2	89.14%
3-9	0.2033	4,805,343	5	3.93	93.90%	0.2359	0.3758	0.6065	10	3	51.56%
10-99	0.4392	10,378,260	29	21.50	94.00%	0.0846	0.1597	0.2935	10	6	21.91%
100-999	0.2146	5,070,981	231	180.00	94.10%	0.0286	0.0561	0.0953	10	18	7.69%
+1000	0.0511	1,207,342	2,051	1660.00	94.40%	0.0085	0.0190	0.0313	10	56	2.60%

Notes: Beginning-of-quarter employment is defined as all jobs held by a worker at the same establishment during the quarter and during the previous quarter. Statistics are computed across all state-year-quarters within a table. The "All" category of establishments includes private as well as state and local government but excludes federal employment. All tables include all valid QWI age groups with the exception of any table including education, in which case only jobs with workers age 25 and older are included. For statistic definitions for beginning of quarter employment, please see their respective equations in the accompanying text: Count 6, Total Variation 15, Missingness Ratio 16, Coefficient of Variation 32. (a) Undefined value.

Table 3: Summary of Total Variability of All Full-Quarter Employment (*EmpS*) by Table and Count

Table and *EmpS* count range	Proportion of Cells	Number of Cells	Median Count	Median Total Variation	Median Rubin Missingness Rate (Percent)	Quantiles of Coefficient of Variation			Median Approximate 90% Confidence Intervals Margin of Error		
						5th	Median	95th	Median df	Count	Percent
All (Private plus State and Local)											
Age x Gender											
100-999	0.0001	3	961	402.00	79.30%	0.0209	0.0209	0.0209	14	27	2.81%
+1000	0.9999	44,941	56,533	4060.00	32.10%	0.0003	0.0011	0.0035	87	82	0.14%
Race x Ethnicity											
zero measured value, after rounding	0.0002	7	9	5.16	95.20%	0.2022	0.2589	0.4127	9	3	35.81%
10-99	0.0323	1,088	48	35.15	95.70%	0.0737	0.1267	0.2891	9	8	17.53%
100-999	0.1687	5,685	455	299.00	94.60%	0.0122	0.0420	0.0848	10	24	5.76%
+1000	0.7989	26,928	11,454	3550.00	81.90%	0.0001	0.0039	0.0235	13	80	0.52%
Gender x Education											
+1000	1.0000	22,472	143,578	134000.00	96.60%	0.0012	0.0029	0.0081	9	506	0.39%
Industry x County											
zero measured value, after rounding	0.0085	26,395	0	0.27	95.50%	(a)	(a)	(a)	9	1	(a)
1-2	0.0001	445	2	0.20	0.00%	0.1178	0.2565	0.8139	9999	1	32.88%
3-9	0.0273	84,593	7	0.41	0.00%	0.0557	0.1030	0.3814	9999	1	13.20%
10-99	0.2858	884,129	44	5.06	51.60%	0.0228	0.0539	0.1654	33	3	7.05%
100-999	0.4368	1,351,160	287	55.60	75.30%	0.0101	0.0245	0.0595	15	10	3.29%
+1000	0.2414	746,888	2,812	690.00	79.00%	0.0024	0.0081	0.0198	14	35	1.08%
Age x Gender x Industry x County											
zero measured value, after rounding	0.2313	10,443,994	0	0.20	95.80%	(a)	(a)	(a)	9	1	(a)
1-2	0.0052	234,881	2	0.36	74.00%	0.1383	0.3166	0.7228	16	1	42.33%
3-9	0.2281	10,301,188	5	0.75	67.40%	0.0789	0.1672	0.3849	19	1	22.20%
10-99	0.3679	16,614,512	26	4.96	74.30%	0.0352	0.0799	0.1806	16	3	10.68%
100-999	0.1410	6,367,152	220	48.10	77.50%	0.0130	0.0288	0.0605	15	9	3.86%
+1000	0.0265	1,197,767	1,914	417.00	76.90%	0.0040	0.0097	0.0191	15	27	1.30%
Race x Ethnicity x Industry x County											
zero measured value, after rounding	0.6294	21,431,041	0	0.19	96.00%	(a)	(a)	(a)	9	1	(a)
1-2	0.0053	179,385	2	0.66	92.60%	0.2632	0.6042	0.8922	10	1	82.90%
3-9	0.1210	4,119,278	5	2.08	89.70%	0.1232	0.2998	0.5754	11	2	40.88%
10-99	0.1417	4,825,719	26	8.94	86.00%	0.0393	0.1088	0.2576	12	4	14.76%
100-999	0.0746	2,541,058	249	67.80	81.50%	0.0126	0.0305	0.0701	13	11	4.11%
+1000	0.0281	955,663	2,460	637.00	80.80%	0.0031	0.0091	0.0202	13	34	1.22%
Gender x Education x Industry x County											
zero measured value, after rounding	0.0989	2,281,075	0	0.26	95.60%	(a)	(a)	(a)	9	1	(a)
1-2	0.0053	121,268	2	1.37	94.10%	0.4295	0.6500	0.9421	10	2	89.19%
3-9	0.2129	4,907,073	5	3.90	93.90%	0.2359	0.3763	0.6074	10	3	51.64%
10-99	0.4341	10,007,726	28	21.10	93.90%	0.0849	0.1608	0.2945	10	6	22.06%
100-999	0.2026	4,671,711	229	178.00	94.10%	0.0286	0.0562	0.0953	10	18	7.71%
+1000	0.0462	1,065,581	2,021	1620.00	94.30%	0.0087	0.0190	0.0312	10	55	2.60%

Notes: Total employment is defined as all jobs held by a worker at the same establishment during the quarter. Statistics are computed across all state-year-quarters within a table. The "All" category of establishments includes private as well as state and local government but excludes federal employment. All tables include all valid QWI age groups with the exception of any table including education, in which case only jobs with workers age 25 and older are included. For statistic definitions for total employment, please see their respective equations in the accompanying text: Count 6, Total Variation 15, Missingness Ratio 16, Coefficient of Variation 32. (a) Undefined value.

Table 4: Summary of Total Variability of All Total Payroll (*Payroll*) by Table and Count

Table and *EmpTotal* count range	Proportion of Cells	Number of Cells	Median Payroll	Median Total Variation	Median Rubin Missingness Rate (Percent)	Quantiles of Coefficient of Variation			Median Approximate 90% Confidence Intervals Margin of Error		
						5th	Median	95th	Median df	Count	Percent
All (Private plus State and Local)											
Age x Gender											
+1000	1.0000	46,480	431,844,381.50	4.06E+11	30.00%	0.0004	0.0014	0.0078	99	822,066.74	0.18%
Race x Ethnicity											
10-99	0.0181	632	248,224.00	2.23E+09	97.70%	0.1058	0.2015	0.4318	9	65,310.59	27.87%
100-999	0.1223	4,263	2,153,721.00	1.82E+10	96.30%	0.0348	0.0672	0.1469	9	186,580.78	9.29%
+1000	0.8596	29,965	80,813,010.00	4.83E+11	83.80%	0.0004	0.0063	0.0449	12	942,546.65	0.85%
Gender x Education											
+1000	1.0000	23,240	1,344,933,652.50	2.35E+13	96.60%	0.0016	0.0038	0.0110	9	6,704,480.56	0.53%
Industry x County											
zero measured value, after rounding	0.0024	8,225	0.00	9.59E+06	99.80%	0.0529	0.4142	1.2418	9	4,282.93	57.28%
1-2	0.0886	309,518	47,962.00	1.43E+07	0.00%	0.0000	0.0656	0.5918	9999	4,846.55	8.41%
3-9	0.0120	41,946	27,741.50	6.22E+06	0.00%	0.0299	0.0854	0.5657	9999	3,196.39	10.95%
10-99	0.2126	743,122	223,252.00	1.72E+08	56.10%	0.0193	0.0590	0.2425	28	17,213.63	7.74%
100-999	0.4137	1,445,825	1,652,146.00	2.85E+09	83.00%	0.0100	0.0314	0.0903	13	72,079.42	4.25%
+1000	0.2708	946,236	19,043,488.50	5.83E+10	82.90%	0.0033	0.0111	0.0315	13	326,004.16	1.49%
Age x Gender x Industry x County											
zero measured value, after rounding	0.1426	7,973,123	0.00	5.66E+05	99.90%	0.0000	0.3606	5.6209	9	1,040.49	49.87%
1-2	0.1515	8,471,709	4,432.00	4.34E+05	52.60%	0.0000	0.1250	0.9335	32	862.07	16.35%
3-9	0.1846	10,324,414	17,514.00	1.00E+07	87.70%	0.0458	0.1819	0.6249	11	4,311.55	24.80%
10-99	0.3423	19,140,564	116,677.00	1.23E+08	87.70%	0.0309	0.0949	0.2717	11	15,121.17	12.94%
100-999	0.1481	8,279,653	1,215,134.00	2.00E+09	87.90%	0.0131	0.0368	0.0921	11	60,974.46	5.01%
+1000	0.0309	1,728,489	14,814,716.00	3.50E+10	85.30%	0.0045	0.0126	0.0311	12	253,725.03	1.71%
Race x Ethnicity x Industry x County											
zero measured value, after rounding	0.4663	19,553,448	0.00	3.51E+06	100.00%	0.0387	0.4623	2.8196	9	2,591.10	63.93%
1-2	0.1778	7,456,341	5,093.00	8.21E+06	98.60%	0.0824	0.6106	1.2213	9	3,962.81	84.45%
3-9	0.1158	4,856,222	21,430.50	6.60E+07	97.00%	0.1083	0.4033	0.8508	9	11,235.78	55.77%
10-99	0.1368	5,735,093	129,598.00	3.71E+08	94.20%	0.0441	0.1494	0.3933	10	26,430.12	20.49%
100-999	0.0733	3,073,969	1,398,948.00	3.58E+09	89.90%	0.0141	0.0421	0.1103	11	81,578.26	5.74%
+1000	0.0301	1,262,815	16,919,665.00	5.45E+10	86.00%	0.0039	0.0127	0.0339	12	316,612.12	1.72%
Gender x Education x Industry x County											
zero measured value, after rounding	0.0639	1,787,333	240.00	2.85E+05	99.60%	0.0000	0.2135	1.9833	9	738.34	29.53%
1-2	0.1360	3,803,163	7,036.00	2.13E+07	98.90%	0.3025	0.6337	1.2283	9	6,382.94	87.65%
3-9	0.1649	4,610,815	25,593.00	1.41E+08	98.10%	0.2569	0.4754	0.8347	9	16,422.56	65.75%
10-99	0.3847	10,755,591	165,697.00	1.13E+09	97.50%	0.1005	0.2040	0.4132	9	46,491.16	28.21%
100-999	0.2001	5,593,317	1,541,163.00	1.33E+10	97.30%	0.0356	0.0740	0.1405	9	159,498.65	10.23%
+1000	0.0504	1,407,901	17,357,576.00	2.23E+11	96.80%	0.0112	0.0266	0.0535	9	653,105.94	3.67%

Notes: Total Payroll is defined only over total employment. It is calculated by summing the earnings for the reference quarter for total employment. See the table on total employment for the relevent counts. Statistics are computed across all state-year-quarters within a table. The "All" category of establishments includes private as well as state, and local government but excludes federal employment. All tables include all valid QWI age groups with the exception of any table including education, in which case only jobs with workers age 25 and older are included. For statistic definitions for beginning of quarter employment, please see their respective equations in the accompanying text: Total payroll 28, Total Variation 31, Missingness Ratio 16, Coefficient of Variation 32. (a) Undefined value.

Table 5: Summary of Total Variability of All Average Monthly Earnings (*EarnS*) by Table and Count

Table and *EmpS* count range	Proportion of Cells	Number of Cells	Median Average Monthly Earnings	Median Total Variation	Median Rubin Missingness Rate (Percent)	Quantiles of Coefficient of Variation			Median Approximate 90% Confidence Interval Margin of Error		
						5th	Median	95th	Median df	Count	Percent
All (Private plus State and Local)											
Age x Gender											
100-999	0.0001	3	1,779.00	15,000.00	87.00%	0.0688	0.0688	0.0688	11	166.99	9.38%
+1000	0.9999	44,941	2,176.00	5.39	23.50%	0.0004	0.0012	0.0062	164	2.99	0.15%
Race x Ethnicity											
3-9	0.0002	7	2,335.00	307,000.00	97.10%	0.1820	0.3357	0.6261	9	766.30	46.42%
10-99	0.0323	1,088	2,127.50	68,000.00	96.20%	0.0616	0.1201	0.3266	9	360.65	16.62%
100-999	0.1687	5,685	2,225.00	8,170.00	94.70%	0.0133	0.0406	0.0979	10	124.03	5.57%
+1000	0.7989	26,928	2,508.50	140.00	75.30%	0.0004	0.0046	0.0302	15	15.86	0.61%
Gender x Education											
+1000	1.0000	22,472	2,844.00	63.60	94.70%	0.0011	0.0028	0.0080	10	10.94	0.38%
Industry x County											
zero measured value, after rounding	0.0013	4,351	0.00	2,490,000.00	99.50%	(a)	(a)	(a)	9	2182.38	(a)
1-2	0.0859	288,667	2,135.00	6,710.00	0.00%	0.0000	0.0520	0.3027	9999	104.98	6.66%
3-9	0.0252	84,593	1,520.00	9,030.00	0.00%	0.0196	0.0665	0.2977	9999	121.79	8.53%
10-99	0.2632	884,129	1,969.00	6,060.00	44.00%	0.0147	0.0405	0.1292	46	101.22	5.26%
100-999	0.4022	1,351,160	2,264.00	1,810.00	71.30%	0.0071	0.0197	0.0531	17	56.73	2.62%
+1000	0.2223	746,888	2,722.00	337.00	71.50%	0.0022	0.0071	0.0201	17	24.48	0.95%
Age x Gender x Industry x County											
zero measured value, after rounding	0.0016	69,616	0.00	2,790,000.00	99.70%	(a)	(a)	(a)	9	2310.11	(a)
1-2	0.2095	9,158,213	1,276.00	9,220.00	12.10%	0.0000	0.0852	0.4918	611	123.19	10.93%
3-9	0.2357	10,301,188	1,469.00	18,700.00	73.50%	0.0298	0.0971	0.3096	16	182.80	12.97%
10-99	0.3801	16,614,512	1,868.00	8,520.00	76.90%	0.0197	0.0532	0.1441	15	123.74	7.13%
100-999	0.1457	6,367,152	2,383.00	2,040.00	77.40%	0.0079	0.0207	0.0526	15	60.55	2.77%
+1000	0.0274	1,197,767	3,152.00	481.00	73.30%	0.0029	0.0075	0.0191	16	29.32	1.01%
Race x Ethnicity x Industry x County											
zero measured value, after rounding	0.0027	51,636	0.00	6,010,000.00	99.90%	(a)	(a)	(a)	9	3390.54	(a)
1-2	0.3472	6,643,546	1,892.00	252,000.00	97.50%	0.0570	0.2777	0.7523	9	694.27	38.40%
3-9	0.2153	4,119,278	2,009.00	132,000.00	93.90%	0.0580	0.1859	0.4701	10	498.54	25.51%
10-99	0.2522	4,825,719	2,145.00	25,100.00	87.90%	0.0248	0.0748	0.2077	11	216.01	10.20%
100-999	0.1328	2,541,058	2,324.00	2,840.00	80.90%	0.0087	0.0238	0.0624	13	71.95	3.21%
+1000	0.0499	955,663	2,763.00	430.00	75.90%	0.0027	0.0079	0.0212	15	27.80	1.06%
Gender x Education x Industry x County											
zero measured value, after rounding	0.0021	53,160	0.00	3,550,000.00	98.80%	(a)	(a)	(a)	9	2605.83	(a)
1-2	0.1652	4,096,117	1,803.00	451,000.00	98.20%	0.1473	0.3915	0.8583	9	928.79	54.15%
3-9	0.1979	4,907,073	1,899.00	237,000.00	96.20%	0.1280	0.2628	0.5388	9	673.29	36.35%
10-99	0.4035	10,007,726	2,205.00	57,100.00	94.70%	0.0499	0.1110	0.2480	10	327.89	15.22%
100-999	0.1884	4,671,711	2,580.00	10,400.00	94.60%	0.0187	0.0411	0.0878	10	139.94	5.63%
+1000	0.0430	1,065,581	3,188.00	2,370.00	94.40%	0.0066	0.0160	0.0374	10	66.80	2.20%

Notes: Average Monthly Earnings is defined only over full-quarter jobs. It is calculated by taking the earnings for the reference quarter for full-quarter jobs and dividing by 3. See the table on full-quarter employment for the relevent counts. Statistics are computed across all state-year-quarters within a table. The "All" category of establishments includes private as well as state, and local government but excludes federal employment. All tables include all valid QWI age groups with the exception of any table including education, in which case only jobs with workers age 25 and older are included. For statistic definitions for beginning of quarter employment, please see their respective equations in the accompanying text: Average Monthly Earnings 20, Total Variation 23, Missingness Ratio 16, Coefficient of Variation 32. (a) Undefined value.

4.1 Interpretation of the Tables

Tables 1-5 have the same structure.[13] The major row label is the level of QWI tabulation. For example, the row labeled "Age × Gender" refers to the collection of tabulations stratified by year, quarter, ownership (private), state, age category, and gender. The published QWI data conform to the schema listed here: `http://lehd.ces.census.gov/doc/QWIPU_Data_Schema.pdf`. Refer to this page for categories of the stratifying variables. The minor row label characterizes the publication cell by its size. For Table 2 the size classes are based on the values of beginning-of-quarter employment. For Tables 1 and 4 the size classes are based total employment, and for Tables 1 and 5, the classes are based on full-quarter employment. The complete set of size classes we summarize is:

- Zero measured value, after rounding, which means that the estimated value is zero.
- 1-2, which means that the estimated value is in the interval [1,2] after rounding.
- 3-9, which means that the estimated value is in the interval [3,9] after rounding.
- 10-99, which means that the estimated value is in the interval [10,99] after rounding.
- 100-999, which means that the estimated value is in the interval [100,999] after rounding.
- +1000, which means that the estimated value is in the interval [1000,max] after rounding.

The column labeled "Proportion of Cells" shows the proportion of all cells in the major row category that lie within the minor row category size class. For example, the value 1.000 in Table 1, for the Age × Gender publication tables in the +1000 size class indicates that all the cells in the Year × Quarter × Ownership (all) × State × Age category × Gender publication tables have at least 1,000 employees in the cell for the publication period 1990:1 through 2012:1. The column labeled "Number of Cells" gives the total number of cells published for this major row category in the indicated count range. Using the

[13] Appendix Tables A.6 to A.10 also follow this structure.

same row as an example, the value 46,480 means that there are this many unique cells in the Year × Quarter × Ownership (all) × State × Age category × Gender publication tables for the same period.

For Tables 1, 2, and 3 the next column is "Median Count," which is the median value of the tabulation variable *EmpTotal*, (respectively, *Emp*, *EmpS*) in the cells covered by that row. Using the same example row in Table 1, the value 91,515 is the median value of total employment in the 46,480 Age × Gender cells summarized in that row. For Table 5, the next column is "Median Average Monthly Earnings," which is the median value of average monthly earnings for all of cells tabulated in a row of the table. For Table 4, the next column is "Median Payroll." For all five tables, we report medians rather than averages for most statistics. We compute all tabulations over all tabulated cells used for that row. Upon careful review of the summary tables, we found outlier cells to have undo influence on summary statistics based on averages. We therefore use medians, which believe best summarizes the "typical" cell for a given stratification.

For Tables 1-5, the next column "Median Total Variation" reports the median value of the Rubin total variation for the cells tabulated in that row. In Tables 1, 2, and 3 this is the median value of tvc_k^* from equation 15 variable *EmpTotal* (respectively, *Emp*, *EmpS*). In Table 4 this is the median value of tvw_k^* from equation 31, and from Table 5 it is the median value of tvz_k^* from equation 23. The values tabulated in this column are the overall summary measures of data quality for the five released total quality measures.

For Tables 1-5, the next column "Median Rubin Missingness Rate (Percent)" reports the median value of the missingness ratio stated as a percentage. The reported statistic is the median value in a cell over all cells used in the indicated row. See sub-section 4.3 for a discussion of the interpretation of this data quality statistic.

Again for Tables 1-5, the next four columns report the "Quantiles of the Coefficient of Variation, where the coefficient of variation is defined in equation 32. These columns restate the square root of the Rubin total variation statistic as a ratio to the estimated value

of the publication statistic. These statistics on the coefficient of variation can be used to assess the proportionate total variation around the published value arising from all sources of error. See the discussion in sub-section 4.2.

The final three columns of Tables 1-5, "Approximate median 90% Confidence Interval Margin of Error" report the Rubin approximate degrees of freedom and the margins of error of the median 90% approximate confidence intervals. The "margin of error" is one-half of the 90% confidence interval width. For $EmpTotal$, Emp, and $EmpS$, we compute the approximate degrees of freedom using the moment-matching formula from Rubin and Schenker (1986)

$$df_k^* = (L-1)\left(1 + \frac{L}{L+1}\frac{\bar{vc}_k^*}{bc_k^*}\right)^2 \tag{33}$$

where the appropriate within-variance (equation 13) and between-variance (equation 11) is used in the numerator and denominator, respectively. To compute the approximate degrees of freedom for confidence intervals for $EarnS$, we use the within-variance from equation 22 and the between variance from equation 21 in equation 33. In all cases, $L = 10$. The same logic applies to $Payroll$ with its corresponding equations. The margin of error for the count is computed by multiplying the square root of the average total variance by the t-statistic value for probability 0.05 with the degrees of freedom indicated in the "df" column. The margin of error for the percent is calculated by multiplying the average coefficient of variation by the same t-statistic, then expressing the result as a percentage.

The engaged reader may notice a seeming anomaly when viewing the summary median degrees of freedom in Tables 1-5. The median degrees of freedom for the Industry × County, employment sizes 3-9 row, reside at our imposed upper bound and appear curious compared to the other rows. This is especially true compared to the row above. The Industry × County, employment sizes 1-2 row has a much smaller median degrees of freedom, in line with the other rows in the summary tables. Upon further inspection, this is not an error. The apparent anomaly lies with the suppression rules in the QWI public-use tables and the preponderance of multi-unit employers in a given cell. To understand

the role of the multi-unit employers, recall that county and industry are singly imputed on the establishment-level employer characteristics file that is the source data for these two workplace characteristics. The only source of between variance at the Industry × County level is through the imputation of a workplace to a worker – called the unit-to-worker impute in the technical documentation, which can result in variance in the industry and geography codes associated with a particular job. Cells with employment in the range 3-9, have few employer firms, and the distribution of firms skews towards single-establishment firms. Single-unit firms have no unit-to-worker imputation, and are not a source of between variance. The predominance of single-unit firms in these cells pushes the degrees of freedom towards its upper bound. The other important factor is the suppression of most cells in the public-use data that contain estimated employment counts of 1-2. In the cell counts in Tables 1-5 one sees a sharp dip in the cell count. This is not a representative sub-population of cells, which leads to anomalous looking summary results. When one looks at Table 4, *Payroll*, for which items are never suppressed, one sees that the median degrees of freedom is also at the upper bound, which is what we would expect given the small employment size and the predominance of single-unit employers in these cells.

We interpret the approximate median 90% confidence interval margins of error for the counts as providing evidence about the overall reliability of counts of $EmpTotal$, Emp, and $EmpS$ for cells that lie in the indicated count range. For example, the margin of error for the count associated with the Age × Gender cell in Table 2, +1000 row is 94, and the average value of Emp in that row is 70,233. The approximate 90% confidence intervals are 70,233 +/- 94. The approximate confidence interval margins of error for counts are most useful for assessing the reliability of estimates in the range zero (after rounding) to nine, although we provide them for all count ranges.

We interpret the approximate average 90% confidence intervals stated in percentages as providing evidence on the relative reliability of counts of $EmpTotal$, Emp, and $EmpS$. Using the same row as an example, we have the relative 90% confidence interval of 70,233

+/- 0.13%. The approximate confidence interval margins of error stated in percentages are useful for assessing the reliability of estimates in the range 10 to 1,000 and over – that is, for the cells containing the vast bulk of employment.

4.2 Computing Confidence Bounds for Published Estimates of *EmpTotal*, *Emp*, *EmpS*, *Payroll*, and *EarnS*

In this subsection, we explain how to use the distribution files to compute more accurate 90% confidence intervals for published QWI and LODES data.[14] The distribution files contain total variation measures computed using equation 15 for *EmpTotal*, *Emp*, and *EmpS*, and equation 23 for *EarnS*, and equation 31 for *Payroll*. The components of the confidence intervals used to compute the results in Tables 1-5 can be replaced by the comparable quantities in the distribution files to improve the accuracy of the confidence intervals.

Find the appropriate distribution table (corresponding to a major row label in Tables 1-5) and the appropriate rows of the distribution file (corresponding to the desired values of the stratifying variables). Take the square root of the total variation measure to form confidence intervals for the reported values of *EmpTotal*, *Emp*, and *EmpS*, *Payroll*, and *EarnS*. Divide the square root of the total variation measure by the level of the published value to form percentage confidence intervals. Derive the within variance using the total variance and the missingness ratio as

$$\bar{vc}_k^* = (1 - mrc_k^*)\, tvc_k^* \,, \tag{34}$$

where the appropriate value of the missingness ratio and the total variance should be used for the different statistics, respectively. Derive the between variance using total variance,

[14] Found here: `http://doi.org/10.3886/E100590V1`.

within variance and the total number of implicates according to the formula

$$bc_k^* = \frac{L}{L+1}\left(tvc_k^* - \bar{vc}_k^*\right) \ , \tag{35}$$

where $L = 10$. Finally, compute the approximate degrees of freedom according to equation 33.

To form a more accurate confidence interval for the level of the published indicator, multiply the square root of the total variance for that measure by the appropriate value from the t-distribution with the degrees of freedom indicated by equation 33 and the desired confidence level. To form a more accurate confidence interval for the percentage variation of the published indicator, divide the margin of error calculated for the level by the value of the published statistic. We recommend using confidence intervals calculated from employment counts for cells with tabulations from zero to nine. We recommend using confidence intervals calculated from the percentage variation in employment for cells with tabulations of 10 or more. For confidence intervals on average monthly earnings of full-quarter employment, we recommend using percentage variation.

Users of LODES/OTM can use Table 2 to estimate approximate confidence intervals for workplace employment counts published in OTM or calculated directly from LODES. Find the major row label in Table 2 that most closely approximates the stratification used in the LODES/OTM workplace summary. Generally, that will be one of the tables with detailed "county-level" geographic stratification combined with demographic or firm-level variables. There is no QWI equivalent for the earnings category stratification available in LODES. Once the closest suitable QWI table has been selected, select the row with the count range that corresponds to the employment count for which a confidence interval is desired. For employment counts of zero to nine, use the count margin of error to form an approximate 90% confidence interval. For employment counts of 10 or more, use the percentage margin of error to form an approximate 90% relative confidence interval. If other

levels of confidence are required, use the degrees of freedom estimate in the same row to look up the correct t-statistic for the desired confidence level, then compute count margins of error using the square root of the average total variation in the row or compute percentage margins of error using the average coefficient of variation in the row.

4.3 Discussion of the Interpretation of Missingness Ratios and Data Quality

The Rubin total variance measure is the appropriate statistic to summarize the total quality of the published indicators for total employment, beginning-quarter employment, full-quarter employment, total payroll, and average monthly earnings of full-quarter employees. It is clear from Tables 1-5 that total variation declines monotonically, in percentage terms, as the number of jobs in the tabulation value increases. This is hardly surprising, but careful attention to the magnitudes of these percentage total variations (in the coefficient of variation columns) shows that for even the most detailed tables and for the stratifiers associated with the largest missingness ratios, the tabulations are very reliable when based on job counts of at least 10, and moderately reliable for job counts of three to nine. This conclusion remains valid even if the very pessimistic assessment of total variation (the 95th percentile of the distribution of the coefficient of variation) is used.

The missingness ratio, therefore, is not a measure of total quality. Instead, it is an indicator of which components of the infrastructure used to compute the QWI and LODES can be most improved by investments in data that reduce the amount of edit and imputation required to estimate that component.

Two components stand out in this regard: education in comparison with worker age and gender. Education is imputed for the vast majority (about 87%) of the individuals in the LEHD infrastructure based on a multistage ignorable missing data model. By contrast, worker age and gender are imputed for less than seven percent of the individuals. And race and ethnicity are imputed for about 18% of the individuals. Looking closely at

the average coefficients of variation for the Age × Gender × Industry × County table in comparison with the Gender × Education × Industry × County table, we see that for every count range, the Age × Gender table has less total variation than the Gender × Education table. The explanation is that the missingness ratio never falls below 91% for the Gender × Education table, whereas it varies between 41% and 71% for most of the Age × Gender table. The statistics confirm that the quality of the Gender × Education table can only be improved by reducing the contribution from missing data. The analysis also confirms that even with very large missingness ratios, the Gender × Education tabulations have acceptable total variation for tabulations involving at least 10 employees.

5 Conclusion

We have conducted the first comprehensive total error and variability analysis of five major publication variables in the Quarterly Workforce Indicators, namely the two key employment indicators and the most widely used earnings indicator. The beginning-of-quarter employment variable from QWI is also the primary tabulation variable in the LEHD Origin-Destination Employment Statistics; hence, our analysis is also applicable to workplace tabulations directly from LODES or displayed in OnTheMap. Our analysis reveals that the very smallest tabulations (estimated zeros and counts of one or two) are not particularly reliable in the sense that they could easily range from zero to three. Tabulations of three to nine are more reliable in the sense that the 90% confidence bound is generally less than plus or minus four. Tabulations involving 10 or more jobs are very reliable having percentage variation that declines from a worst case of plus or minus 31% (count range 10-99, tables involving education) to a best case of plus or minus less than one percent (count range +1000, tables involving firm age).

To the best of our knowledge, no other widely used statistical system based on administrative records has produced a comprehensive total error analysis to which the results

in this paper can be compared. As compared to survey-based estimates like those derived from the American Community Survey, for example, the QWI employment and earnings tabulations have accuracy comparable to the accuracy of the ACS (U.S. Census Bureau, 2015) even when comparing state and PUMA-level estimates in the ACS to county and core-based statistical areas in the QWI. The LODES/OTM estimates for sub-county geographies and small sub-populations have much lower total error than estimates from the ACS from comparably-sized sub-populations. Designed surveys like the ACS deliver statistics on a much broader set of variables, and can be used for analyses that are far outside the scope of the QWIs or LODES/OTM. But our analyses demonstrate that the total error of an administrative-records based publishing system that combines data from many sources can compare favorably with much more expensive survey-based systems.

References

Abowd, J. M., K. Gittings, K. L. McKinney, B. E. Stephens, L. Vilhuber, and S. Woodcock (2012). Dynamically consistent noise infusion and partially synthetic data as confidentiality protection measures for related time-series. In *Federal Committee on Statistical Methodology, 2012 Research Conference Papers*. Office of Management and Budget.

Abowd, J. M., B. E. Stephens, L. Vilhuber, F. Andersson, K. L. McKinney, M. Roemer, and S. Woodcock (2009). The LEHD Infrastructure Files and the Creation of the Quarterly Workforce Indicators. In T. Dunne, J. B. Jensen, and M. J. Roberts (Eds.), *Producer Dyanamics: New Evidence from Micro Data*, pp. 149–230. University of Chicago Press.

Abowd, J. M. and L. Vilhuber (2005). The Sensitivity of Economic Statistics to Coding Errors in Personal Identifiers. *Journal of Business & Economic Statistics 23*(2), 133–152.

Benedetto, G., J. Haltiwanger, J. Lane, and K. Mckinney (2007). Using Worker Flows to Measure Firm Dynamics. *Journal of Business & Economic Statistics 25*(3), 299–313.

Biemer, P. P. (2010). Total survey error: Design, implementation, and evaluation. *Public Opinion Quarterly 74*(5), 817.

Bishop, Y. M., S. E. Fienberg, and P. W. Holland (1975). *Discrete Multivariate Analysis: Theory and Practice*. Cambridge, MA: MIT Press.

Cochran, W. G. (1977). *Sampling Techniques* (3rd ed.). Wiley.

Dinur, I. and K. Nissim (2003). Revealing information while preserving privacy. *Proceedings of the twenty-second ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, 202–210.

Dwork, C., F. McSherry, K. Nissim, and A. Smith (2006). Calibrating Noise to Sensitivity in Private Data Analysis. *Tcc*, 265–284.

Eurostat (2014). Quality report of the European Union Labour Force Survey 2013.

Evans, T., L. Zayatz, and J. Slanta (1998). Using Noise for Disclosure Limitation of Establishment Tabular Data. *Journal of Official Statistics 14*(4), 537–551.

Groves, R. M., F. J. J. Fowler, M. P. Couper, J. M. Lepkowski, E. Singer, and R. Tourangeau (2004). *Survey Methodology*. Hoboken, NJ: John Wiley & Sons.

Groves, R. M. and L. Lyberg (2010). Total survey error: Past, present, and future. *Public Opinion Quarterly 74*(5), 849–879.

Haney, S., A. Machanavajjhala, J. M. Abowd, M. Graham, M. Kutzbach, and L. Vilhuber (2017). Utility cost of formal privacy for releasing national employer-employee statistics. In *Proceedings of the 2017 International Conference on Management of Data*, Volume forthcoming of *SIGMOD '17*. ACM.

Horrigan, M., P. Phipps, and S. Fricker (2014). Development of a Quality Framework and Quality Indicators at the Bureau of Labor Statistics. In *Joint Statistical Meetings 2014 -Government Statistics Section*, pp. 325–338.

Li, Q. and J. Racine (2003). Nonparametric Estimation of Distributions with Categorical and Continuous Data. *Journal of Multivariate Analysis 86*(2), 266–292.

Little, R. J. and D. B. Rubin (2002). *Statistical Analysis with Missing Data* (2nd ed.). Hoboken, NJ: Wiley.

Mulry, M. H. and A. D. Keller (2017). Comparison of 2010 Census Nonresponse Follow-Up Proxy Responses with Administrative Records Using Census Coverage Measurement Results. *Journal of Official Statistics 33*(2), 455–475.

Reid, G., F. Zabala, and A. Holmberg (2017). Extending TSE to Administrative Data: A Quality Framework and Case Studies from Stats NZ. *Journal of Official Statistics 33*(2), 477–511.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley.

Rubin, D. B. and N. Schenker (1986). Multiple Imputation for Interval Estimation From Simple Random Samples With Ignorable Nonresponse. *Journal of the American Statistical Association 81*(394), 366–374.

Starsinic, M. (2011). Incorporating a Finite Population Correction Factor into American Community Survey Estimates. In *JSM Proceedings, Survey Research Methods Section*, Alexandria, Virginia, pp. 3621–3631. American Statistical Association.

U.S. Census Bureau (2003a). 2000 Census of Population and Housing, Public Use Microdata Sample, United States: Technical Documentaiton.

U.S. Census Bureau (2003b). PUMS Accuracy of the Data, The American Community Survey.

U.S. Census Bureau (2015). American Community Survey Multiyear Accuracy of the Data.

Zhang, X., M. L. King, and R. J. Hyndman (2006). A Bayesian Approach to Bandwidth Selection for Multivariate Kernel Density Estimation. *Computational Statistics & Data Analysis 50*(11), 3009 – 3031.

Appendices

A Details of the Methodology for Imputing Missing Birth date, Sex, Race, Ethnicity, and Education[15]

The LEHD data come from state UI systems' reports of a worker, a firm, and the worker's quarterly earnings. The data the Census Bureau receives from the states contain no information on worker characteristics including age, sex, race, ethnicity, and education. These individual characteristics are a unique attribute of the QWI and LODES. In order to provide the individual characteristics, the Census Bureau attaches its own surveys as well as administrative data from other U.S. government agencies to the LEHD UI data. In cases where the outside surveys and administrative data are not sufficient to account for all characteristics for all workers, the characteristics are imputed.

This appendix documents the methodology for imputing missing individual characteristics in the LEHD infrastructure files. The appendix describes the data sources for the individual characteristics that form the basis of the imputation. The candidate imputation models and the basis for their selection are also documented. After explaining the monotone missing data pattern and the final implementation of the imputation process, the quality of the imputation is assessed. At the end of the process, the complete set of individual characteristics is stored in the Individual Characteristics File (ICF), which stores the individual characteristics for all workers who appear in the LEHD UI data including 10 draws of the imputation model for each characteristic that is imputed.

The main source data for race and ethnicity is the 2000 Decennial Census of Population and Housing (short form). For birth date and sex, the Census Numident – the Census Bureau's version of the Social Security Administration's (SSA) Social Security Number

[15]Portions of this appendix are based on an unpublished technical memo dated February 1, 2011 by John Abowd, Henry Hyatt, Mark Kutzbach, Erika McEntarfer, Kevin McKinney, Michael Strain, Lars Vilhuber, and Chen Zhao.

(SSN) master database – is the only source. In cases where the race and ethnicity data are incomplete (i.e., an individual's response to the 2000 Census or ACS was not available) an imputation of an individual's race and ethnicity category was computed conditional on the limited race and ethnicity information available in the Census Numident file (if available). The source data for education is the 2000 Decennial Census of Population and Housing Sample Data (long form). Since education is dynamic, particularly for young workers, education data are only imputed for workers aged 25 and older.[16]

The missing characteristics are imputed using a Bayesian version of the continuous-discrete multivariate product kernel density (KDE) approach. In some instances a multinomial model with Dirichlet priors was employed. These missing data follow a monotone pattern. The characteristics are imputed in three stages, with data completed from the previous stage used in the imputation model for the next stage. The end results is 10 implicates of completed data drawn from estimates of the posterior distribution of the characteristics.

To assess the out-of-sample performance of the imputation model, two separate tests are used. First, the completed race, ethnicity, and education variables were matched to a sample of respondents from the ACS (2000-2010). These comparisons show highly accurate imputation rates, particularly for the larger race and ethnicity groups: White (95% accuracy), Black (90% accuracy), Asian (85% accuracy), and Hispanic (80% accuracy). For education, the results are adequate, but they do not display the same level of accuracy.

In addition to conducting ACS comparisons, the geographic variability captured by our education model was also assessed. Using a sub-sample of workers who have a recorded 2000 Decennial Census (long form) education response, tabulations of beginning-of-quarter employment, full-quarter employment, and average quarterly wages for full-quarter em-

[16]The current version of the ICF at press time includes the ACS as a data source for education as well as race and ethnicity. When the research team fit the imputation models assessed in this appendix, the ACS was not used as a source of individual characteristics, which is what made it suitable for assessing model fit.

ployees by both the actual and imputed value are calculated. These comparisons show close correspondence, particularly for wages. At the statewide level, the difference between full-quarter wages within education categories for reported and imputed education ranges from -6.8% to +8.0% with some cells within 0.2%. The share of beginning of quarter employment in each education category varies by a range of -5.3 to 6.6 percentage points with most cells within 2 percentage points.

The rest of this appendix proceeds as follows. Section A.1 describes the selection of the missing data model for imputing the individual characteristics, Section A.2 details the implementation of the models for each of the characteristics, and Section A.3 assesses the quality of the imputation.

A.1 Methodological Approach

Missing, birth date, sex, race, ethnicity, and education were imputed using multiple imputation following Rubin (1987). The candidate imputation models were implemented and tested before selecting a final procedure at each stage of the imputation. We compared several different estimators: (i) the standard Li and Racine (2003) mixed continuous-discrete KDE (LR); (ii) a Bayesian Li-Racine method based on an approach developed by Zhang et al. (2006) for estimating the posterior of the bandwidth parameter (ZH); (iii) a multinomial distribution with a Dirichlet prior combined with Bayesian bootstrap re-sampling (BB); (iv) a cold deck (the equivalent of hot deck methods when all the data are given) (CD); and (v) a naïve method (modal imputations in sub groups) (NA).

To assess the performance of each candidate, a 3-dimensional distribution for birth year, race/ethnicity, and education was created using data from the Current Population Survey (CPS).[17] Using balanced half-sample cross validation, the research question examined was: with 100% imputation rates, what are the Kullback-Leibler Divergence (KL) and Mean Squared Error (MSE) losses associated with each of these methods, assuming

[17]Specifically, the 1998 through 2005 pooled March data.

ignorable missing data.

The combined years of the CPS were treated as a synthetic population of 170,000 individuals. For each of the candidates the KL and MSE criteria were estimated using the CPS data. The KL was computed by comparing the actual and imputed distributions. Half-samples were created randomly by assigning in-scope individuals permanently to A and B sub-populations of equal sizes. All models were fit on sub-population A, then used to impute sub-population B, subsequently the process was reversed with the estimates based on the B sample used to impute A. Hence, every member of the population received imputed values for every model based on an out-of-estimation-sample forecast. All the estimators were compared for a variety of stratifying schemes. KL and MSE performances were considered when adopting strategies for choosing stratifiers used in the final implementation.

The ZH and LR methods underestimated the KL and MSE losses, using BB as the standard, but often by less than 10%. In many cases, the ZH and LR methods were effectively indistinguishable from the BB. LR, ZH and BB substantially out-performed both the cold-deck and naïve models. Up to two levels of stratifiers, with a total of eight sub-populations, were tested.[18] There were large (one or two orders of magnitude) improvements in the KL and MSE loss estimates as stratifiers were added. The BB, ZH, LR, and CD methods all led to the same conclusions about which stratifiers to consider first, and to the conclusion that with subpopulations of 20,000 from a population of 170,000, all stratifiers improved the KL and MSE measurably. The NA model performed poorly, which was expected. The BB, ZH, and LR models all outperformed the CD, and were roughly comparable.

LR and ZH methods were implemented for birth date, sex, race and ethnicity, and partly for education. A variant of BB was also implemented for education. The two KDE methods perform well relative to BB, directly handle continuous data, and allow greater

[18]This approximately evenly stratified the CPS population into sub-populations of about 20,000 records each.

flexibility in the actual implementation. Occasionally the cells created by the stratifiers became too small to estimate with the KDE methods necessitating the use of BB.

A.2 Implementation

The missing data follow a special monotone pattern, allowing us to complete the data in three stages. Birth date, sex and place of birth (completed but not used in any tabulations) have the least missing data (about 5% of cases), and are (almost) always missing if race, ethnicity or education are missing. Race and ethnicity are missing for about 18% of the individuals, and are always missing if education is missing. The variables with the fewest missing data values (sex, birth date, and place of birth) were imputed first. Missing race and ethnicity were imputed next, taking the imputed values for birth date, sex, and place of birth as given. Finally, missing education was imputed.[19]

At each stage, the variables imputed in the previous stage(s) along with various detailed work history, firm, and co-worker characteristics derived from the unemployment insurance wage data were used to create cells. The design of this stratification scheme was based on the tests described above using the CPS test synthetic population.

The models were fit using persons with complete information at each stage with a full set of interacted explanatory variables. Intuitively, the models partition observations by stratifying variables (workers) into cells, and then estimate the distribution of interest for each cell. For example, a model for education would estimate the education distribution for a cell of white women ages 35-44 with non-missing education. Observations who are white women ages 35-44 and who are missing education would then receive 10 draws from the distribution fit on that cell.

[19]The monotone missing data pattern is a result of the process by which SSNs are attached to the 2000 Decennial. Sex, date of birth, and place of birth are available on the Census Numident. These data are virtually complete because they are necessary for the administration of the program. Only valid SSNs can be attached to a given 2000 Decennial record, generating the monotone missing data pattern.

A.2.1 Birth date, Sex, and Place of Birth

The Social Security Administrations Numident is the source for birth date and sex. The Numident is the Social Security Administrations master file of issued SSNs, which contains a near universe of birth date and sex information of U.S. workers. Approximately 97% of workers in the LEHD data can be matched to the Numident. Birth date and sex are multiply imputed for approximately 7% of records.

A non-parametric KDE is used to estimate the joint distribution of sex and age conditional on various observed characteristics. The model is state specific, and uses the complete set of yearly earnings and employment indicator variables spanning the entire time a states records are available. The estimated model parameters are used to calculate a predicted probability the record is male. Age is imputed is a similar manner. QWI and LODES report age in eight discrete categories. For the purpose of imputing birth date, a record with missing birth date information is assigned into one of the eight age categories using the KDE model similar to the sex imputation. Date of birth is then assigned based on the distribution of ages within each of the eight age categories for entering workers. As with sex, 10 independent draws assign 10 separate dates of birth for each record contain missing date of birth.

The sex and place of birth variables are unordered categorical, and age is real numeric. For estimating the distributions, the following stratifiers were used:

- Modal place of birth non-native-born coworkers
- Proportion of coworkers that are male ($> 50\%$).
- New worker indicator.

A.2.2 Race and Ethnicity

To implement the race and ethnicity imputation, the following steps were taken. First, since the 2000 Census Short Form provided substantial respondent flexibility for reporting race and ethnicity, it was necessary to simplify the reporting for the imputation models.

The vast majority of respondents chose single race and ethnicity categories. A small fraction of the population (less than 3%) reported multiple race and/or ethnicity responses. In compliance with OMB statistical policy, the multiple race responses were collapsed into a single category (two or more races), and ethnicity was collapsed to two responses (Hispanic and not Hispanic). For the respondents who reported "some other race," the actual response was set to missing and they were imputed into one of the OMB-approved race categories.

The non-parametric unordered KDE modeled the joint distribution of race and ethnicity. The model incorporates the imputed age and sex information from the previous step. The race variable is grouped into seven different categories, and the ethnicity variable into just two: Hispanic and not Hispanic. The principal source for race and ethnicity information is the Census 2000 short form. Subsequent iterations of the model also incorporate race and ethnicity information from the American Community Survey. Approximately 82% of persons found in the LEHD have valid race and ethnicity information from either the Census 2000 or American Community Survey data. For the remaining records with missing race or ethnicity, the values are multiply imputed.

The ethnicity categories on the QWI tabulations by race and ethnicity are:

1. Hispanic or Latino
2. Not Hispanic or Latino

The race categories on the QWI tabulations by race and ethnicity are:

1. White Alone
2. Black or African American Alone
3. Asian Alone
4. Native Hawaiian or Other Pacific Islander Alone
5. American Indian or Alaska Native Alone
6. Two or More Races.

Race and Ethnicity are both unordered categorical variables. The stratifiers for stage B include both age and place of birth from stage A. In addition, there are:

- Collapsed race/ethnicity cells from the Census Numident
- Average yearly earnings quartiles.
- Coworker fraction white and coworker fraction Hispanic.
- Co-resident fraction white and co-resident fraction Hispanic.

A.2.3 Education

The data for the education imputation come from the 2000 Decennial Census Long Form. Approximately 7% of LEHD workers have valid education information.[20] The modal response "high school graduate, no college" was retained exactly. Three additional categories were created by collapsing the other responses from the 2000 Decennial Census Long Form education variable. The education categories are:

1. Less than a high school diploma
2. High school graduate, no college
3. Some college or Associates degree
4. Bachelor's degree or above.

Unlike race and ethnicity, which were modeled as time-invariant, a person is at risk to accrue additional formal education after entering the workforce, however, this risk declines with age. Individuals generally complete high school before age 20, while Bachelor's degrees are disproportionately attained between the ages of 22 and 25. To ameliorate concerns of younger workers attending post-secondary education, the QWI and LODES only report and impute education data for workers at least age 25.

[20]The subsequent inclusion of the ACS after 2000 increases the number of workers with valid education information to 15%.

A Bachelor's degree is almost always required to pursue a graduate degree. Associate degree and some college were collapsed into a single category. The resulting ordered categorical education variable allows the use of an informative kernel when estimating the education density. The stage C stratifiers include the imputed variables from stages A and B as well as:

- Place of birth by income quantile.
- Native and Non-native status.
- Modal NAICS (6 categories) for dominant job.
- Collapsed race and ethnicity cells.
- Coworker fraction male.
- Full-quarter earnings deciles.
- Co-resident fraction white and co-resident fraction Hispanic.

For education, the multinomial-Dirichlet (called BB above, but with no final bootstrap step) was used. Although the LR KDE has improved out of sample performance for imputing education, in the current implementation a fully interacted log-linear model with flat priors was used instead because of its superior performance in small geographic cells. When using stratifiers with a large number of outcomes (detailed geography in particular), the number of cells became too large relative to the sample size. To solve this problem we estimated a log-linear model with a reduced set of parameters. This allows us to include stratifiers as main effects only or with limited interactions, improving overall performance. This is essentially a small-area estimator for which the mean vector is estimated by the main effects associated with the stratifiers and local effects are estimated from the log-linear model.

A.3 Quality of the Results

For imputations of race and ethnicity, the chief quality check is a detailed comparison of the completed race and ethnicity variables to a matched sample of respondents on the

American Community Survey (ACS). Because the ACS was not used as an input for the imputation models, the ACS provides an out-of-sample performance assessment.

The primary question posed by this analysis was: how frequently does the missing data model impute individuals with no 2000 Census race or ethnicity information to the same race or ethnicity category they indicate in the ACS? The results show very accurate imputations for most race and ethnicity groups, although there is variation across ACS race and ethnicity categories. The highest levels of accuracy, defined here as imputing a response on the LEHD infrastructure consistent with ACS race/ethnicity response, are for the largest race and ethnicity groups: White (95% accuracy), African-American (90% accuracy), Asian (85% accuracy), and Hispanic (80% accuracy).

Defining an accuracy measure for Native American populations (American Indian, Alaska Native, Native Hawaiian or Pacific Islander) proved more problematic as a matched sample of Census/ACS respondents indicated that a large share of these respondents diverged in their race responses between the Census and the ACS. However, for Native Americans that answer both surveys consistently, imputed LEHD race corresponds to self-reported race well over half of the time. A sizable share of self-reported Native Hawaiians and Pacific Islanders are imputed to Asian in the LEHD infrastructure, in part because a key stratifier for the race imputation (the race variable on the Census Numident) does not separate Pacific Islanders from Asians.

For imputations of education, multiple levels of quality checks were employed. In addition to comparisons with the ACS, a comparison of key QWI variables for three sample states by Education and Education $\times$ Sex was analyzed using both reported education and imputed education. This analysis used a sample of workers in the LEHD infrastructure that has a reported Census 2000 long form education response, for which an imputed response was also generated for this assessment. Beginning-of-quarter employment (B), full-quarter employment (F), and average monthly wages for full-quarter employees (Z_W3) were studied using both respondent-supplied education and imputed education.

These indicators were computed for both the reported value of education and for each of the 10 education implicates. The difference between the value of the QWI indicator using reported education and the average value for the indicator using imputed education over the 10 implicates was studied.

For B, F, and Z_W3 analyzing the Education $\times$ Sex breakdown at the statewide level, the correspondence is quite close. In statewide Education $\times$ Sex tabulations, the difference between average full-quarter wages within categories for reported and imputed education ranges from -8.1% to +9.4%. The share of beginning-of-quarter employment in each education category varies by a range of -5.3 to +6.6 percentage points with the smallest difference being less than 0.001 percentage points at the statewide level. Differences in male/female wage gaps and employment by education across states are largely retained in the imputed results.[21]

A.3.1 ACS Results

To construct the review of imputation quality the results were merged with the ACS. First, three years of person-level data from the ACS were appended together. The same ICF variables used in the imputation were constructed from the unedited responses on the ACS. The education, race, and ethnicity characteristics constructed from the ACS were then merged into the newly created ICF by PIK. Due to the dynamic nature of education, only persons at least 25 years of age after April 1st 2000 (according to ICF variable *dob*1) were retained for the analysis.

The ICF records were then stratified for each variable. The records were partitioned by variable according to whether they contained a corresponding valid ACS response. Records were then further subdivided into whether or not the ICF variable was imputed creating four mutually exclusive and collectively exhaustive groups. The group containing records for which there was no corresponding valid ACS variable, and in which the value

[21]For disclosure limitation, all results are rounded to three significant digits.

was not imputed, serves as the baseline distribution for each variable. For the two groups for which a valid ACS response exits – ICF variable imputed, and not imputed – the distribution of the ICF variable was computed conditional on the ACS response for each of the two groups.

In addition to the conditional distribution means, confidence intervals were computed for each value of the distribution using the Rubin methodology (within- and between-implicate variance) to draw confidence intervals around the each category using all implicates of the imputed data. Standard errors are calculated using the following formula (described in U.S. Census Bureau (2003a)),

$$stnd\_error = D\sqrt{\frac{S-1}{B}\left(acc_{pct}\right)\left(1-acc_{pct}\right)} \tag{A.1}$$

where D is the corresponding US design factor for the standard error, S is the number of persons in each of the mutually exclusive categories corresponding to a particular variable minus 1, and B is the population count over age 25 according to the 2000 Decennial Census SF3 file for each category.

To account for over-sampling of some populations, for persons not imputed in the ICF and not matching to the ACS, variable-specific design factors were taken from the "Accuracy of Microdata Sample Estimates: Census 2000 PUMS Standard Error Design Factors (U.S. Census Bureau, 2003a)." For Persons matching to the ACS, variable-specific design factors were taken from U.S. Census Bureau (2003b).

For each category of each race, ethnicity and education variable, the imputation model was more informative than a random allocation across categories would have been. The models assigned a higher share of individuals to the same category as those persons responded in the ACS than would be expected if the imputation models assigned categories completely at random from the aggregate distribution. The analysis shows, however, that there is considerable variation in imputation quality across variables.

Tables A.1, A.2, and A.3 show the results and 90-10 confidence intervals for the imputation quality analysis for the variables education, ethnicity, and race, respectively. Each table contains the results for each variable broken out by the individual categories of the variable as reported in the ACS. Table A.1 displays the results for education. The major row heading has the categories for the four possible ACS responses as well as the category for ICF records who do not match to a valid education category. The latter group is the first row in the table. The minor row heading for "Not in ACS" indicates that in addition to not matching to the ACS, this group includes only ICF records whose education categories were not imputed. Moving across the first row, the remaining columns give the education distribution for this group. The remaining rows of Table A.1 give the distribution of education conditional on a particular ACS value of education. The minor row headings indicate that these groups are further partitioned by whether the ICF value was imputed.

Figure A.1 depicts the two education distributions for each value of education in the ACS. This depicts graphically what is presented in Table A.1. Each sub-figure corresponds to an ACS value. The blue bars give the distribution of those records, which were not imputed. This serves as the target distribution. The red bar gives the distribution of the records which were imputed. Ideally, this would line up perfectly with the blue bars, but that is not always the case. The green bar shows the overall distribution for records, which were not imputed, and which did not match to the ACS. This is the baseline distribution, and it does not vary across ACS categories. In addition to education, figures depicting impute quality by matching to the ACS are available for race and ethnicity. For each category of each variable, the impute model should not be expected to be much better than the matched ACS responses, so the red line is unlikely to be greater than the green line. The green line does not always equal 1 (or 100%) for the specified ICF category because some people responded differently on the Decennial Census or Numident than they did on the ACS.

The education figures show the most accurate imputations were for the "High School"

Table A.1: Distribution of ICF Categories across ACS Response Categories, Education

Distribution of catagories in ICF	90% CI	< High School	High School	Some College	≥ Bachelor
Not in ACS					
Baseline: not imputed	Upper	13.8%	29.6%	30.5%	26.2%
	Mean	13.7%	29.6%	30.5%	26.2%
	Lower	13.7%	29.6%	30.5%	26.2%
ACS: Less than High School					
Impute: imputed, ACS is '< High School'	Upper	26.3%	33.8%	26.9%	14.3%
	Mean	26.0%	33.5%	26.6%	14.0%
	Lower	25.6%	33.1%	26.4%	13.6%
Target: not imputed, ACS is '< High School'	Upper	80.8%	15.0%	4.2%	1.1%
	Mean	80.4%	14.7%	4.0%	1.0%
	Lower	80.0%	14.3%	3.8%	0.9%
ACS: High School					
Impute: imputed, ACS is 'High School'	Upper	14.8%	35.6%	31.6%	18.8%
	Mean	14.6%	35.4%	31.4%	18.6%
	Lower	14.5%	35.1%	31.2%	18.4%
Target: not imputed, ACS is 'High School'	Upper	6.5%	81.5%	12.0%	0.9%
	Mean	6.3%	81.2%	11.7%	0.8%
	Lower	6.1%	80.8%	11.4%	0.7%
ACS: Some College					
Impute: imputed, ACS is 'Some College'	Upper	11.0%	29.9%	33.5%	26.4%
	Mean	10.8%	29.7%	33.3%	26.2%
	Lower	10.7%	29.5%	33.1%	25.9%
Target: not imputed, ACS is 'Some College'	Upper	1.4%	11.3%	85.2%	3.0%
	Mean	1.3%	11.0%	84.8%	2.9%
	Lower	1.2%	10.7%	84.5%	2.7%
ACS: ≥ Bachelors					
Impute: imputed, ACS is '≥ Bachelors'	Upper	6.1%	19.0%	28.6%	47.2%
	Mean	6.0%	18.8%	28.4%	46.9%
	Lower	5.8%	18.6%	28.1%	46.6%
Target: not imputed, ACS is '≥ Bachelors	Upper	0.3%	0.9%	4.8%	94.6%
	Mean	0.2%	0.8%	4.6%	94.3%
	Lower	0.2%	0.7%	4.4%	94.1%

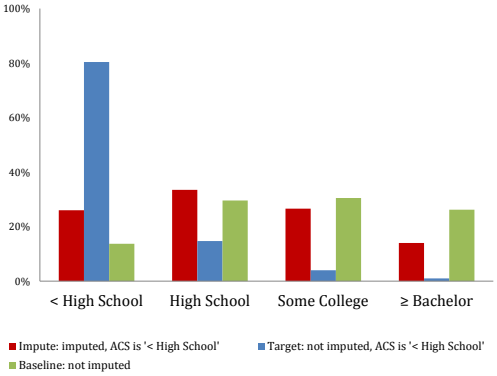
Notes: 90% CI are 90% confidence intervals of the mean. Major row heading is the value of the ACS variable. Minor row heading is the value of the ICF variable. Major row header "Not in ACS" denotes records that did not match to the ACS.

and "Bachelor's degree and above" categories. The blue line in Figure A.1(d) shows that a little over 94% of records reporting "Bachelor's degree and above" in the 2000 Decennial also reported the same value in the ACS. Of the records imputed into the "Bachelor's degree and above" category and matched to the ACS (red bar), slightly less than 47% had the same value in the ACS. The corresponding values for "High School," Figure A.1(b), are 81.2% (blue bar) and 35.4% (red bar).

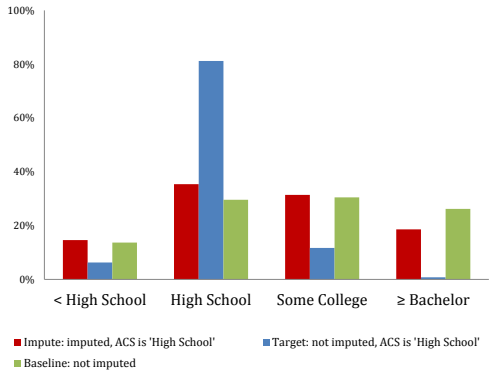
The imputations for the education categories "Less than High School" and "Some College" were somewhat less successful, as measured by correspondence with the ACS. The red bar in Figure A.1(c) gives a rate of 84.8% correspondence between the Decennial and ACS for records which were not imputed and had a value of "Some College." The blue bar depicting correspondence for records which were imputed shows a rate of 33.3%. For "Less than High School" in Figure A.1(a), the two rates are 80.4% (red bar) and 26.0% (blue bar). The lower rate of correspondence for all education values compared to "Bachelor's degree and above" are expected, as some Decennial respondents will have completed more schooling upon responding the ACS at a later date.

For ethnicity, the imputation procedure was more accurate than with education. The population for ethnicity is 90.7% "not Hispanic" versus 9.3% "Hispanic" according to the 2000 Decennial. Figure A.2(a) shows that conditional on reporting "not Hispanic" in the ACS, approximately 94.4% are imputed into the "not Hispanic" group compared to 99.6% of ACS respondents who were not imputed and report being "not Hispanic" in the Decennial Census as well as the ACS. For the Hispanic group, depicted in Figure A.2(b), these numbers are 80.0% and 94.8%, respectively.

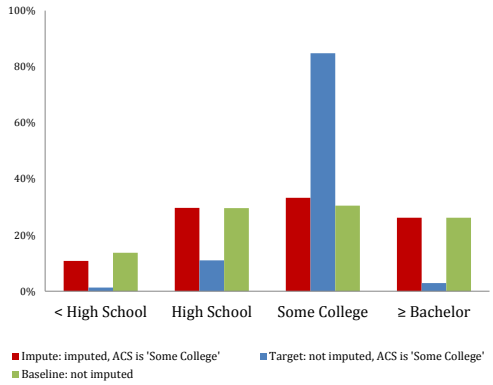
For race, results vary by ACS category. White, Black, and Asian have highly accurate imputations. For these groups, the results are depicted in Figure A.3. For White, Black, and Asian, the rates imputed into those categories conditional on the same ACS response is 94.5%, 89.5%, 83.7%, respectively. This shows relatively high quality as the target distributions are 99.3%, 96.7%, and 94.5%, for White, Black, and Asian, respectively.



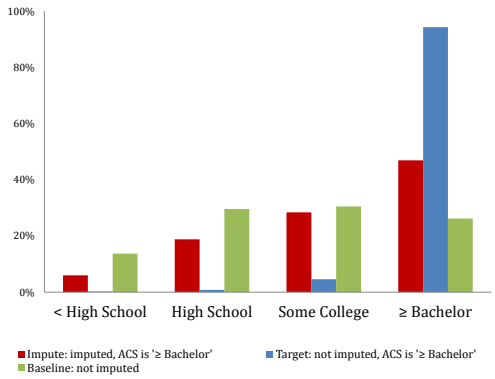
(a) Less than High School



(b) High School



(c) Some College



(d) Bachelor's degree and above

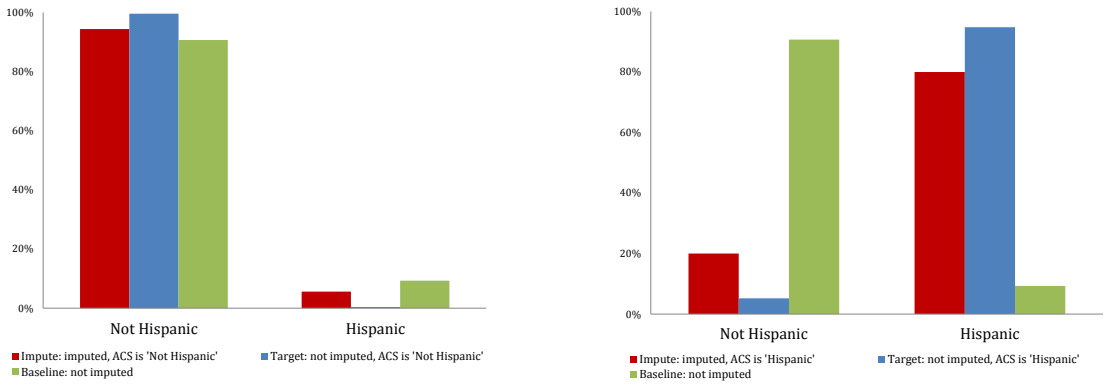
Figure A.1: Impute versus Target: Education

Notes: Sub-figure titles correspond to the value of the education variable in the ICF. The blue bars show the distribution of education in the ICF among records that were not imputed. The red bars shows the distribution of education in the ICF among imputed records. The green bars show the distribution of education among records in the ICF that were not imputed *and* did not match to the ACS. The green bars do not vary across sub-figures. See Table A.1 for more detail.

Table A.2: Distribution of ICF Categories across ACS Response Categories, Ethnicity

Distribution of catagories in ICF	90% CI	Not Hispanic	Hispanic
Not in ACS			
Baseline: not imputed	Upper	90.7%	9.3%
	Mean	90.7%	9.3%
	Lower	90.7%	9.3%
ACS: Not Hispanic			
Impute: imputed, ACS is 'Not Hispanic'	Upper	94.7%	6.0%
	Mean	94.4%	5.6%
	Lower	94.0%	5.3%
Target: not imputed, ACS is 'Not Hispanic'	Upper	99.7%	0.4%
	Mean	99.6%	0.4%
	Lower	99.6%	0.3%
ACS: Hispanic			
Impute: imputed, ACS is 'Hispanic'	Upper	21.7%	81.7%
	Mean	20.0%	80.0%
	Lower	18.3%	78.3%
Target: not imputed, ACS is 'Hispanic'	Upper	5.5%	95.0%
	Mean	5.2%	94.8%
	Lower	5.0%	94.5%

Notes: 90% CI are 90% confidence intervals of the mean. Major row heading is the value of the ACS variable. Minor row heading is the value of the ICF variable. Major row header "Not in ACS" denotes records that did not match to the ACS.



(a) Not Hispanic

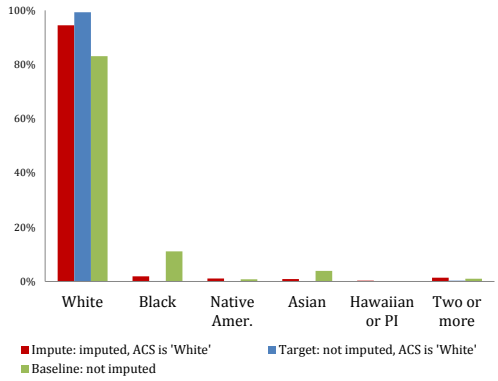
(b) Hispanic

Figure A.2: Impute versus Target: Ethnicity

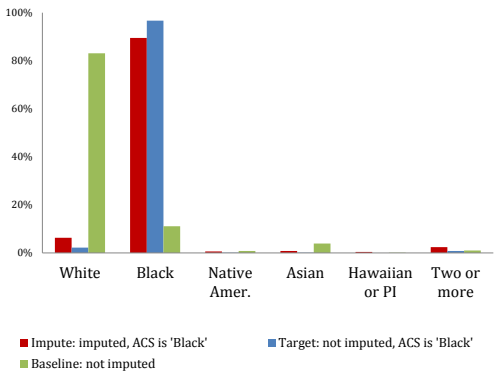
Notes: Sub-figure titles correspond to the value of the ethnicity variable in the ICF. The blue bars show the distribution of ethnicity in the ICF among records that were not imputed. The red bars shows the distribution of ethnicity in the ICF among imputed records. The green bars show the distribution of ethnicity among records in the ICF that were not imputed *and* did not match to the ACS. The green bars do not vary across sub-figures. See Table A.2 for more detail.

For the race categories with much smaller populations, the comparison to the ACS did not yield as accurate imputations. The groups Native American or Alaskan Native, and Hawaiian or Pacific Islander are 0.8% and 0.1%, respectively, of the U.S. population according to the 2000 Census. Conditional on having an ACS response in the same category, 39.2% were imputed into the Native American or Alaskan Native category (Figure A.3(d)), and 8.0% into Hawaiian or Pacific Islander (Figure A.3(e)). This is compared to target shares of 71.3% and 47.0%, respectively. For the Hawaiian or Pacific Islander, the majority of those responding as such on the ACS were imputed into the White and Asian categories at approximately equal rates. For Native American or Alaska Native, Figure A.3(d) shows the majority were imputed into the white category.

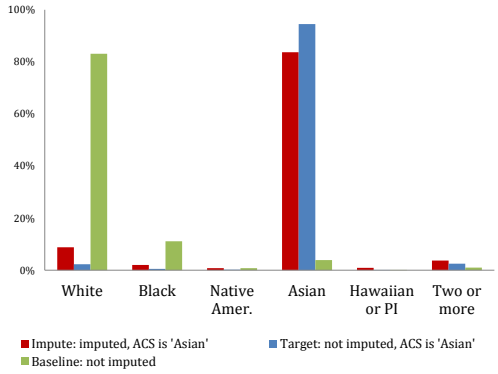
The category "Two or More Races" and "Some Other Race" also have inconsistent responses across input data. Those responding as "Two or More Races" are 1.0% of the population. Their target distribution is 34.5% of ACS respondents who report two or more races and who have the same response in the 2000 Census. For the records imputed from



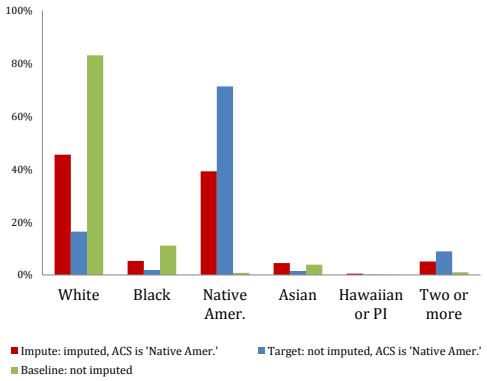
(a) White



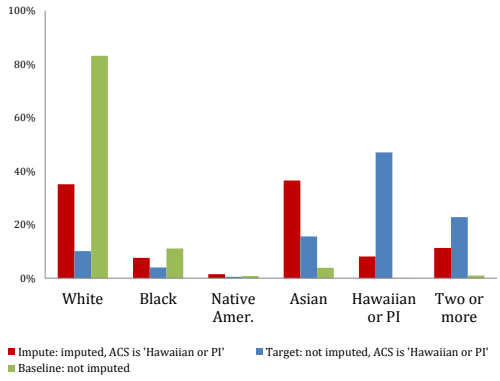
(b) Black



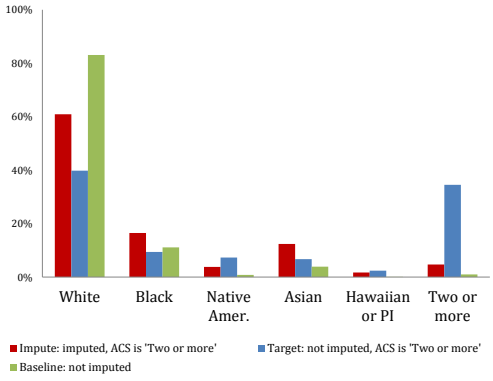
(c) Asian



(d) Native American or Alaska Native



(e) Hawaiian or Pacific Islander



(f) Two or More Races

Figure A.3: Impute versus Target: Race

Notes: Sub-figure titles correspond to the value of the race variable in the ICF. The blue bars show the distribution of race in the ICF among records that were not imputed. The red bars shows the distribution of race in the ICF among imputed records. The green bars show the distribution of race among records in the ICF that were not imputed *and* did not match to the ACS. The green bars do not vary across sub-figures. See Table A.3 for more detail.

the 2000 Census who report two or more races in the ACS, only 4.7% were imputed into the two or more races category. The other records were mostly imputed into the White, Black, and Asian categories as seen in Figure A.3(f). Note that "Some Other Race" is not an imputation category. Respondents to the ACS who answered "Some Other Race" were largely imputed to "White," with a large portion to "Two or More Races."

A.3.2 Comparison to the "D Sample"

The previous section examined the quality of the imputation at the person level. The next set of results asks how the imputation model fairs when used to reproduce LEHD public-use statistics. To do this, a simple comparison of key QWI variables is carried out for three sample states by Education, and Education × Sex, using both reported education and imputed education. This analysis uses the "D sample," a sample of workers in the ICF that have a Census 2000 long form education response. Here we compare beginning-of-quarter employment, full-quarter employment, and average quarterly wages for full-quarter employees using the QWI variables calculated using both respondent education and imputed education. The question of interest posed here is a simple one: for the sample of workers for whom reported education is known, do the QWI statistics show substantially different patterns when imputed education is used to tabulate the statistics rather than respondent education?

For this analysis beginning-of-quarter employment (B), full-quarter employment (F), and wages for full-quarter employees (Z_W3) are computed directly from the internal Employment History File, rather than the production system equivalent, using the standard definitions but not the fuzz factors. These indicators are computed for both the reported value of education and for each of the 10 implicates of the imputed education value. For the sake of simplicity in interpretation, we report the difference between the value of the indicator using reported education compared to the average value for the indicator using imputed education over the 10 implicates. While this is a simplification, as the variation

Table A.3: Distribution of ICF Categories across ACS Response Categories, Race

Distribution of catagories in ICF	90% CI	White	Black	Native Amer.	≥ Asian	Hawaiian & PI	≥ Two or More
Not in ACS							
Baseline: not imputed	Upper	83.1%	11.1%	0.8%	3.9%	0.1%	1.0%
	Mean	83.1%	11.1%	0.8%	3.9%	0.1%	1.0%
	Lower	83.1%	11.1%	0.8%	3.9%	0.1%	1.0%
ACS: White							
Impute: imputed, ACS is 'White'	Upper	94.8%	2.0%	1.2%	1.0%	0.4%	1.5%
	Mean	94.5%	1.9%	1.1%	0.9%	0.3%	1.4%
	Lower	94.2%	1.7%	0.9%	0.7%	0.3%	1.2%
Target: not imputed, ACS is 'White'	Upper	99.3%	0.2%	0.1%	0.1%	0.0%	0.4%
	Mean	99.3%	0.2%	0.1%	0.1%	0.0%	0.3%
	Lower	99.2%	0.2%	0.1%	0.1%	0.0%	0.3%
ACS: Black							
Impute: imputed, ACS is 'Black'	Upper	7.3%	90.8%	1.0%	1.2%	0.6%	3.0%
	Mean	6.3%	89.5%	0.6%	0.8%	0.4%	2.4%
	Lower	5.3%	88.3%	0.3%	0.4%	0.1%	1.7%
Target: not imputed, ACS is 'Black'	Upper	2.4%	96.9%	0.2%	0.2%	0.0%	0.9%
	Mean	2.2%	96.7%	0.1%	0.1%	0.0%	0.8%
	Lower	2.0%	96.5%	0.1%	0.1%	0.0%	0.7%
ACS: Native American							
Impute: imputed, ACS is 'Native Amer.'	Upper	52.9%	8.6%	46.5%	7.6%	1.6%	8.5%
	Mean	45.5%	5.3%	39.2%	4.5%	0.5%	5.1%
	Lower	38.0%	2.0%	31.8%	1.4%	0.0%	1.6%
Target: not imputed, ACS is 'Native Amer.'	Upper	18.1%	2.5%	73.4%	2.1%	0.3%	10.2%
	Mean	16.4%	1.9%	71.3%	1.5%	0.1%	8.9%
	Lower	14.7%	1.2%	69.2%	0.9%	0.0%	7.6%
ACS: Asian							
Impute: imputed, ACS is 'Asian'	Upper	10.7%	3.0%	1.5%	86.2%	1.6%	5.0%
	Mean	8.8%	2.0%	0.8%	83.7%	0.9%	3.7%
	Lower	6.9%	1.1%	0.2%	81.3%	0.3%	2.4%
Target: not imputed, ACS is 'Asian'	Upper	2.6%	0.6%	0.2%	94.9%	0.2%	2.8%
	Mean	2.3%	0.5%	0.2%	94.5%	0.1%	2.5%
	Lower	2.0%	0.3%	0.1%	94.0%	0.1%	2.1%
ACS: Hawaiian & PI							
Impute: imputed, ACS is 'Hawaiian & PI'	Upper	53.1%	17.5%	6.3%	54.7%	18.5%	23.3%
	Mean	35.1%	7.6%	1.5%	36.5%	8.1%	11.3%
	Lower	17.1%	0.0%	0.0%	18.3%	0.0%	0.0%
Target: not imputed, ACS is 'Hawaiian & PI'	Upper	13.7%	6.3%	1.4%	19.9%	53.0%	27.8%
	Mean	10.1%	4.0%	0.5%	15.6%	47.0%	22.8%
	Lower	6.5%	1.6%	0.0%	11.3%	41.1%	17.8%
ACS: Two or More							
Impute: imputed, ACS is 'Two or More'	Upper	89.4%	7.0%	3.3%	3.5%	1.0%	2.5%
	Mean	87.3%	5.6%	2.3%	2.5%	0.6%	1.7%
	Lower	85.2%	4.1%	1.4%	1.6%	0.1%	0.9%
Target: not imputed, ACS is 'Two or More'	Upper	85.1%	6.4%	2.9%	4.6%	0.4%	2.9%
	Mean	84.4%	5.9%	2.6%	4.2%	0.3%	2.6%
	Lower	83.6%	5.5%	2.2%	3.9%	0.2%	2.3%
ACS: Some Other							
Impute: imputed, ACS is 'Some Other'	Upper	64.8%	19.5%	5.3%	15.0%	2.7%	6.4%
	Mean	60.9%	16.5%	3.8%	12.4%	1.7%	4.7%
	Lower	57.0%	13.6%	2.3%	9.8%	0.6%	3.0%
Target: not imputed, ACS is 'Some Other'	Upper	41.0%	10.1%	8.0%	7.3%	2.8%	35.6%
	Mean	39.8%	9.4%	7.3%	6.7%	2.4%	34.5%
	Lower	38.6%	8.6%	6.7%	6.1%	2.0%	33.3%

Notes: 90% CI are 90% confidence intervals of the mean. Major row heading is the value of the ACS variable. Minor row heading is the value of the ICF variable. Major row header "Not in ACS" denotes records that did not match to the ACS.

over the implicates is typically small and generally much smaller than the difference between the average and reported values, it is consistent with the analysis done in the main text of the paper.

Table A.4: Comparison of QWI Variables for the Decennial Sample (D Sample): Actual vs. Imputed Education

	Employment Counts		B Employment Share		F Employment Share		Average Full-quarter wage, (Z_W3)	
Statewide Distribution	B	F	Actual	Imputed*	Actual	Imputed*	Actual	Imputed**
Delaware								
Less than High School	3,510	2,950	10.0%	10.1%	9.7%	9.6%	$6,100	$6,180
High School Graduate	11,400	9,910	32.7%	27.7%	32.4%	27.4%	$7,590	$7,590
Some College or Associates Degree	10,200	8,920	29.3%	30.6%	29.2%	30.6%	$8,950	$9,110
College Graduate or Greater	9,760	8,770	28.0%	31.5%	28.7%	32.4%	$14,900	$13,800
Illinois								
Less than High School	51,500	45,100	8.29%	9.16%	8.07%	8.90%	$6,390	$6,390
High School Graduate	168,000	151,000	27.00%	28.40%	27.00%	28.20%	$7,520	$7,730
Some College or Associates Degree	205,000	185,000	33.00%	31.40%	33.10%	31.50%	$8,880	$9,530
College Graduate or Greater	197,000	178,000	31.80%	31.00%	31.90%	31.40%	$15,700	$15,100
New Jersey								
Less than High School	31,500	27,300	9.14%	8.42%	8.93%	8.12%	$6,860	$6,510
High School Graduate	95,800	85,100	27.80%	21.40%	27.80%	21.10%	$8,550	$8,360
Some College or Associates Degree	93,100	82,600	27.00%	29.20%	27.00%	29.20%	$10,400	$10,500
College Graduate or Greater	125,000	111,000	36.10%	41.00%	36.30%	41.50%	$17,700	$16,400

Notes: *Average share over ten implicates. **Average over ten implicates. Statistics computed for year 2000:2. B denotes beginning-of-quarter employment, and F denotes full-quarter employment.

As can be seen in Table A.4, the comparisons at the state level generally show close correspondence between QWI values using reported education and imputed education, particularly for wages. At the state level, the difference between average full-quarter wages within education categories for reported and imputed education ranges from -6.7% to +8.0% with the smallest difference being less than 0.2%. The share of beginning-of-quarter employment in each education category varies by a range of -4.9 to 6.4 percentage points with the smallest difference being -0.1 percentage points at the statewide level. Overall, differences in the distribution of full-quarter employment between reported and imputed education are similar to those for beginning-of-quarter employment.

For B, F, and Z_W3 for education $\times$ sex at the state level, the correspondence is again quite close. In Table A.5, the difference in education $\times$ sex tabulations between average full-quarter wages within categories for reported and imputed education ranges from -8.1% to +9.4% with the smallest difference being less than 0.09%. The share of

Table A.5: Comparison of QWI Variables for the Decennial Sample (D Sample) by Sex: Actual vs. Imputed Education

		Employment Counts		B Employment Share		F Employment Share		Average Full-quarter wage, (Z_W3)	
Statewide Distribution by Sex		B	F	Actual	Imputed*	Actual	Imputed*	Actual	Imputed**
Delaware									
Female	Less than High School	1,420	1,100	8.45%	8.59%	8.08%	8.04%	$4,470	$4,360
	High School Graduate	5,450	4,750	32.50%	27.10%	32.40%	26.80%	$5,810	$5,610
	Some College or Associates Degree	5,320	4,640	31.80%	32.40%	31.70%	32.40%	$7,120	$7,080
	College Graduate or Greater	4,570	4,080	27.20%	31.90%	27.90%	32.70%	$11,200	$10,600
Male	Less than High School	2,100	1,780	11.50%	11.50%	11.10%	11.00%	$7,190	$7,400
	High School Graduate	5,980	5,170	32.90%	28.30%	32.40%	28.00%	$9,220	$9,320
	Some College or Associates Degree	4,910	4,300	27.00%	29.00%	27.00%	29.00%	$10,900	$11,200
	College Graduate or Greater	5,300	4,700	28.60%	31.20%	29.50%	32.00%	$18,100	$16,900
Illinois									
Female	Less than High School	22,900	20,000	7.46%	8.35%	7.28%	8.11%	$4,500	$4,510
	High School Graduate	81,900	73,700	26.70%	28.90%	26.80%	28.80%	$5,400	$5,490
	Some College or Associates Degree	107,000	95,900	34.90%	33.30%	35.00%	33.40%	$6,600	$6,960
	College Graduate or Greater	95,000	84,900	31.00%	29.40%	30.90%	29.70%	$10,900	$10,800
Male	Less than High School	28,700	25,200	9.10%	9.96%	8.84%	9.66%	$7,880	$7,900
	High School Graduate	85,800	77,200	27.30%	28.00%	27.10%	27.70%	$9,550	$9,990
	Some College or Associates Degree	97,900	88,900	31.10%	29.60%	31.20%	29.60%	$11,300	$12,300
	College Graduate or Greater	102,000	93,400	32.50%	32.50%	32.80%	33.00%	$20,100	$18,900
New Jersey									
Female	Less than High School	14,000	12,100	8.15%	7.80%	7.96%	7.52%	$4,940	$4,730
	High School Graduate	49,300	43,900	28.70%	22.10%	28.80%	21.90%	$6,450	$6,070
	Some College or Associates Degree	49,200	43,600	28.60%	31.10%	28.70%	31.10%	$7,950	$7,890
	College Graduate or Greater	59,500	52,600	34.60%	39.00%	34.60%	39.50%	$12,700	$12,100
Male	Less than High School	17,500	15,200	10.10%	9.03%	9.90%	8.71%	$8,390	$8,040
	High School Graduate	46,500	41,200	26.90%	20.60%	26.80%	20.30%	$10,800	$10,800
	Some College or Associates Degree	43,900	38,900	25.40%	27.40%	25.30%	27.40%	$13,100	$13,600
	College Graduate or Greater	65,100	58,400	37.60%	42.90%	38.00%	43.50%	$22,100	$20,200

Notes: *Average share over ten implicates. **Average over ten implicates. Statistics computed for year 2000:2. B denotes beginning-of-quarter employment, and F denotes full-quarter employment.

beginning-of-quarter employment in each education category varies by a range of -5.3 to 6.6 percentage points with the smallest difference being less than 0.001 percentage points at the statewide level. Differences in male/female wage gaps and employment by education across states are largely retained in the imputed results. Generally and not surprisingly, differences in state comparisons tend to be replicated in smaller cells as well. For instance, the differences in IL between B and F for imputed vs. reported education are very small at the state level and are also very small in the education $\times$ sex cells, while somewhat larger discrepancies in NJ and IL between some education categories are seen in Education $\times$ Sex cells for those two groups.

B Imputation Procedure to Match Research Snapshot and Public-Use Data

The research snapshot used to compute the total variance measures for the QWI differs from the production system used to create the public-use QWI files. The production system does not save the 10 implicates to create the public-use QWI, but these implicates are necessary for the creation of the total variability measures. The research snapshot does not exactly replicate the production QWI statistics due to edits made to each snapshot, which are never reconciled. Due to these edits and rounding, it is sometimes the case that the computed statistics for a given cell do not exactly match. For cells with large employment counts this is a trivial concern as the variance for each statistic is already quite low, and small changes in the magnitude of the statistic result in marginal changes to the coefficient of variation. In cells with small employment counts (less than 10), this is not the case. Small changes in the size of the of employment count lead to large changes in the coefficient of variation. In this appendix we detail how we edit and scale the variance measures to account for the occasional differences in the internal and public-use statistics.

Before proceeding to the edit and scaling algorithm, a brief discussion of the reference

distribution for the coefficient of variation is necessary. The intuition for the edit procedure is that our assumption of equivalent coefficient of variations for the public-use and research snapshots is "reasonable." For any cell with a given employment size, what is reasonable depends on the state, the demographic characteristics and the level of aggregation. We control for these confounding factors by performing the edit procedure separately for each state by ownership type by characteristic crossing. Next, we separate the data by beginning-of-quarter employment; full-quarter employment and average monthly full-quarter earnings; and flow employment and payroll. Within each of the three separate edits, we further separate each cell by its level of aggregation. The edit algorithm is therefore run separately for each state by ownership type by characteristic crossing, by each of the three employment definitions governing the five statistics and by each level of aggregation.

After partitioning the data, the edit algorithm then proceeds as follows. First, we calculate one percent quantiles of the internally calculated employment statistic from the minimum to the maximum. We collapse bins where the employment count is the same for consecutive quantiles leaving us with at most 100 bins for the internally calculated employment statistic. For each bin we calculate the 5th and 95th percentile of the coefficient of variation for the employment statistic as well as average monthly earnings and payroll for full-quarter employment and total employment, respectively. In addition, for each of the five QWI statistics we calculate the median within and between variance as well as the median statistic in the bin.

Once the bins are set, for each record we analyze the bin associated with each of the three public-use employment statistics. If the coefficient of variation for the internally calculated statistic falls either below the 5th percentile or above the 95th percentile of the coefficient of variation in that bin, we use the median within-variance and the median between-variance for that statistic and rescale them accordingly. We then make the total variance, missingness ratio, and degrees of freedom calculations from our edited within-

and between-variance. An example will elucidate the procedure.

Suppose we have a cell with an internally calculated flow employment (M) count of 5 and due to edits and rounding the public-use statistic ($EmpTotal$) is 7. As is typical for low-levels of aggregation and small employment counts, the bins consist of only cells with the same employment counts. That is, the bins consist only of cells with counts of 5, 6, 7, etc. Our public-use flow employment total is 7, so we look at the bin of cells with flow employment counts of 7 and compare our internally calculated coefficient of variation to the distribution in that bin. The coefficient of variation for this cell was calculated from an internal count of 5 and the coefficient of variation is in this example greater than the 95th percentile in the cell. We therefore assign the public-use statistic the median within- and between-variance from the bin, and we scale the two variances by the median flow employment count in the bin, which in this example is simply 7. In this example the median flow employment count in the bin is the same as the public-use statistic negating any change in the variance from scaling, but we use a more reasonable estimate of the within- and between-variance.

C Handling Structural and Sampling Zeros

The public-use QWI files are sparse. If a given cell does not have at least one dollar from a UI-covered job, the cell does not appear in the released data. However, just because a cell does not appear in a particular quarter does not mean that it will not appear in a subsequent quarter. If a cell contains zeros in a given quarter for some combinations of stratifiers but not others, then there are firms operating in that cell, and the zeros are sampling zeros. If there is no evidence of any firm activity in that cell–meaning all combinations of stratifiers show zero employment, then those zeros are all structural zeros. We supplement the unemployment insurance records used as the core inputs to the QWI with firm reports from the QCEW. The QCEW are a firm-level virtual-census of employ-

ment and wages comprising the universe of firms covered by state unemployment insurance systems and some federal employment. The universe of firm activity in the QWI and the QCEW is quite similar but it does not perfectly overlap. To infer firm activity in a given state, year, quarter, county, and NAICS Sector cell, which is the correct frame for distinguishing sampling from structural zeros, we use the union of firm activity from the QWI and QCEW universes. If a cell does not appear in the unemployment insurance microdata, but we find evidence of firm activity – any positive employment in any month or positive wages – from the QCEW we add that cell to the public use file, including all lower levels of aggregation. We flag all sampling zeros with the variable "sample_zero." The five QWI statistics for all sampling zeros are set to zero, and we impute each of their variability statistics.

We impute the variability measures for sampling zeros by exploiting the edit procedure in Appendix B. Recall that in the edit procedure we calculate various moments of the coefficient of variation, within-variance, and between-variance distributions by bins of the internally calculated employment size. The bins are calculated separately for each state, ownership type, characteristic crossing, and aggregation level. We use the median within- and between-variance from the zero bin as the sample zero within- and between- variance. In cases where the aggregation level is too high so as no zero bin exists, we drop down to the next lowest level of aggregation where a zero bin is available and calculate the ratio of the coefficient of variation for the one and zero bins. We scale the within- and between-variance at our reference level of aggregation using the one bin and the ratio calculated from the lower level of aggregation. To summarize, the median within- and between- variance from the zero bin of the edit procedure are used as our imputation of the within-and between variance for sampling zeros. We then derive the total variance, missingness ratio, and degrees of freedom estimates from the within- and between-variance.

D Data Notes & Additional Tables

- North Carolina, Colorado, and Massachusetts are not in the R2012Q4 QWI release and have not been included in the variability files.
- 720 records from the Georgia age by sex all employment file, 588 records from the Georgia race by ethnicity all employment file, and 420 records from the Georgia sex by education all employment file include the NAICS sector 99. This is an error in the release, and these records have been removed from their respective variability files.

Table A.6: Summary of Total Variability of Private Total Employment (*EmpTotal*) by Table and Count

Table and *EmpTotal* count range	Proportion of Cells	Number of Cells	Median Count	Median Total Variation	Median Rubin Missingness Rate (Percent)	Quantiles of Coefficient of Variation			Median Approximate 90% Confidence Intervals Margin of Error		
						5th	Median	95th	Median df	Count	Percent
Private											
Age x Gender											
+1000	1.0000	46,480	80,832	8250.00	43.60%	0.0004	0.0011	0.0036	47	118	0.14%
Race x Ethnicity											
10-99	0.0199	695	53	49.80	96.10%	0.0836	0.1409	0.2605	9	10	19.48%
100-999	0.1306	4,553	452	411.00	95.30%	0.0242	0.0469	0.0935	9	28	6.49%
+1000	0.8495	29,612	13,614	5970.00	86.40%	0.0002	0.0044	0.0274	12	105	0.59%
Gender x Education											
+1000	1.0000	23,240	157,493	192000.00	96.60%	0.0013	0.0031	0.0086	9	606	0.43%
Industry x County											
zero measured value, after rounding	0.0045	12,955	0	0.30	94.40%	(a)	(a)	(a)	10	1	(a)
1-2	0.0001	188	1	0.42	79.35%	0.1982	0.3913	0.9553	14	1	52.73%
3-9	0.0158	45,741	7	0.50	0.00%	0.0614	0.1073	0.3800	9999	1	13.76%
10-99	0.2603	753,131	46	4.82	15.40%	0.0242	0.0511	0.1619	380	3	6.56%
100-999	0.4402	1,273,670	294	53.70	66.30%	0.0105	0.0235	0.0591	20	10	3.12%
+1000	0.2792	807,706	3,058	822.00	76.60%	0.0023	0.0081	0.0201	15	38	1.09%
Age x Gender x Industry x County											
zero measured value, after rounding	0.2019	8,888,449	0	0.21	95.10%	(a)	(a)	(a)	9	1	(a)
1-2	0.0050	220,862	2	0.35	67.20%	0.1564	0.3066	0.7280	19	1	40.71%
3-9	0.2171	9,557,988	5	0.81	61.50%	0.0885	0.1722	0.3887	23	1	22.73%
10-99	0.3796	16,713,425	27	5.35	69.70%	0.0382	0.0815	0.1800	18	3	10.84%
100-999	0.1622	7,142,409	225	55.20	75.40%	0.0142	0.0303	0.0618	15	10	4.06%
+1000	0.0343	1,511,324	1,972	502.00	75.70%	0.0041	0.0103	0.0201	15	30	1.38%
Race x Ethnicity x Industry x County											
zero measured value, after rounding	0.5839	19,047,330	0	0.20	95.20%	(a)	(a)	(a)	9	1	(a)
1-2	0.0058	190,628	2	0.69	91.70%	0.2636	0.6229	0.9257	10	1	85.47%
3-9	0.1330	4,339,494	5	2.35	88.90%	0.1325	0.3167	0.5963	11	2	43.18%
10-99	0.1607	5,240,825	26	10.10	85.30%	0.0426	0.1162	0.2704	12	4	15.76%
100-999	0.0830	2,707,617	249	75.90	79.90%	0.0137	0.0322	0.0745	14	12	4.33%
+1000	0.0336	1,094,612	2,586	766.00	79.20%	0.0031	0.0095	0.0212	14	37	1.28%
Gender x Education x Industry x County											
zero measured value, after rounding	0.0996	2,207,640	0	0.26	94.80%	(a)	(a)	(a)	10	1	(a)
1-2	0.0050	111,105	2	1.35	93.10%	0.4281	0.6538	0.9466	10	2	89.72%
3-9	0.2024	4,484,091	5	3.99	92.70%	0.2395	0.3791	0.6106	10	3	52.03%
10-99	0.4264	9,446,881	29	22.00	92.80%	0.0865	0.1624	0.2961	10	6	22.28%
100-999	0.2119	4,695,684	235	190.00	93.10%	0.0292	0.0569	0.0967	10	19	7.81%
+1000	0.0546	1,209,869	2,089	1790.00	93.50%	0.0087	0.0193	0.0321	10	58	2.65%

Notes: Total employment is defined as all jobs held by a worker at the same establishment during the quarter. Statistics are computed across all state-year-quarters within a table. The "Private" category of establishments includes only private establishments. All tables include all valid QWI age groups with the exception of any table including education, in which case only jobs with workers age 25 and older are included. For statistic definitions for total employment, please see their respective equations in the accompanying text: Count 6, Total Variation 15, Missingness Ratio 16, Coefficient of Variation 32. (a) Undefined value.

Table A.7: Summary of Total Variability of Private Beginning-of-Quarter Employment (*Emp*) by Table and Count

Table and *Emp* count range	Proportion of Cells	Number of Cells	Median Count	Median Total Variation	Median Rubin Missingness Rate (Percent)	Quantiles of Coefficient of Variation			Median Approximate 90% Confidence Intervals Margin of Error		
						5th	Median	95th	Median df	Count	Percent
Private											
Age x Gender											
+1000	1.0000	45,712	61,308	5000.00	37.50%	0.0003	0.0011	0.0035	64	92	0.14%
Race x Ethnicity											
3-9	0.0000	1	9	3.57	92.60%	0.2099	0.2099	0.2099	10	3	28.81%
10-99	0.0282	968	47	36.80	95.90%	0.0783	0.1282	0.2729	9	8	17.73%
100-999	0.1607	5,509	466	328.00	94.70%	0.0115	0.0427	0.0828	10	25	5.86%
+1000	0.8111	27,806	11,716	4000.00	83.20%	0.0002	0.0041	0.0241	13	85	0.55%
Gender x Education											
+1000	1.0000	22,856	132,437	136000.00	96.60%	0.0013	0.0031	0.0085	9	510	0.42%
Industry x County											
zero measured value, after rounding	0.0096	27,314	0	0.29	95.50%	(a)	(a)	(a)	9	1	(a)
1-2	0.0001	313	2	0.37	66.90%	0.1425	0.3279	0.9000	20	1	43.45%
3-9	0.0242	69,274	7	0.42	0.00%	0.0561	0.1009	0.3648	9999	1	12.93%
10-99	0.2916	833,370	44	4.28	21.20%	0.0227	0.0499	0.1609	201	3	6.42%
100-999	0.4286	1,225,043	283	51.40	71.10%	0.0102	0.0235	0.0586	17	10	3.14%
+1000	0.2459	702,856	2,936	747.00	78.70%	0.0024	0.0080	0.0199	14	37	1.08%
Age x Gender x Industry x County											
zero measured value, after rounding	0.2368	10,146,295	0	0.20	95.60%	(a)	(a)	(a)	9	1	(a)
1-2	0.0051	217,085	2	0.33	69.70%	0.1409	0.2971	0.7099	18	1	39.52%
3-9	0.2243	9,610,779	5	0.72	63.10%	0.0811	0.1641	0.3815	22	1	21.69%
10-99	0.3624	15,526,526	26	4.98	72.80%	0.0365	0.0797	0.1786	17	3	10.63%
100-999	0.1432	6,136,088	222	51.50	77.80%	0.0135	0.0296	0.0610	14	10	3.97%
+1000	0.0282	1,210,018	1,931	452.00	77.40%	0.0042	0.0101	0.0196	15	29	1.35%
Race x Ethnicity x Industry x County											
zero measured value, after rounding	0.6229	20,152,114	0	0.19	95.70%	(a)	(a)	(a)	9	1	(a)
1-2	0.0052	168,667	2	0.66	92.30%	0.2579	0.6042	0.8972	10	1	82.90%
3-9	0.1222	3,951,621	5	2.16	89.60%	0.1241	0.3040	0.5810	11	2	41.45%
10-99	0.1465	4,740,254	26	9.16	85.80%	0.0395	0.1103	0.2619	12	4	14.96%
100-999	0.0746	2,411,633	246	69.90	81.40%	0.0130	0.0310	0.0716	13	11	4.18%
+1000	0.0287	926,867	2,513	687.00	80.90%	0.0031	0.0093	0.0205	13	35	1.25%
Gender x Education x Industry x County											
zero measured value, after rounding	0.1166	2,517,116	0	0.26	95.40%	(a)	(a)	(a)	9	1	(a)
1-2	0.0055	118,679	2	1.34	93.80%	0.4257	0.6500	0.9392	10	2	89.19%
3-9	0.2150	4,638,908	5	3.87	93.50%	0.2365	0.3763	0.6062	10	3	51.64%
10-99	0.4195	9,052,346	28	21.00	93.60%	0.0857	0.1620	0.2946	10	6	22.23%
100-999	0.1959	4,228,095	232	183.00	93.90%	0.0288	0.0563	0.0957	10	19	7.73%
+1000	0.0475	1,025,888	2,045	1670.00	94.20%	0.0086	0.0191	0.0315	10	56	2.62%

Notes: Beginning-of-quarter employment is defined as all jobs held by a worker at the same establishment during the quarter and during the previous quarter. Statistics are computed across all state-year-quarters within a table. The "Private" category of establishments includes only private establishments. All tables include all valid QWI age groups with the exception of any table including education, in which case only jobs with workers age 25 and older are included. For statistic definitions for beginning of quarter employment, please see their respective equations in the accompanying text: Count 6, Total Variation 15, Missingness Ratio 16, Coefficient of Variation 32. (a) Undefined value.

Table A.8: Summary of Total Variability of Private Full-Quarter Employment (*EmpS*) by Table and Count

Table and *EmpS* count range	Proportion of Cells	Number of Cells	Median Count	Median Total Variation	Median Rubin Missingness Rate (Percent)	Quantiles of Coefficient of Variation			Median Approximate 90% Confidence Intervals Margin of Error		
						5th	Median	95th	Median df	Count	Percent
Private											
Age x Gender											
100-999	0.0001	6	965	414.00	79.70%	0.0211	0.0211	0.0211	14	27	2.84%
+1000	0.9999	44,938	50,251	3810.00	33.10%	0.0004	0.0012	0.0038	82	80	0.15%
Race x Ethnicity											
3-9	0.0005	17	9	8.14	96.60%	0.1746	0.3394	0.4180	9	4	46.94%
10-99	0.0351	1,184	46	32.70	95.10%	0.0747	0.1279	0.2935	9	8	17.69%
100-999	0.1780	6,001	452	301.00	94.40%	0.0133	0.0420	0.0856	10	24	5.76%
+1000	0.7863	26,506	10,312	3290.00	80.60%	0.0002	0.0042	0.0239	13	77	0.56%
Gender x Education											
+1000	1.0000	22,472	115,661	114000.00	96.40%	0.0014	0.0031	0.0088	9	467	0.43%
Industry x County											
zero measured value, after rounding	0.0142	40,036	0	0.28	95.60%	(a)	(a)	(a)	9	1	(a)
1-2	0.0002	505	2	0.19	0.00%	0.1308	0.2518	0.8485	9999	1	32.27%
3-9	0.0327	92,014	7	0.40	0.00%	0.0571	0.1024	0.3636	9999	1	13.12%
10-99	0.3147	886,839	43	4.21	23.50%	0.0231	0.0509	0.1608	162	3	6.55%
100-999	0.4159	1,172,234	276	51.60	71.80%	0.0105	0.0241	0.0589	17	10	3.22%
+1000	0.2224	626,901	2,870	723.00	78.40%	0.0024	0.0081	0.0200	14	36	1.10%
Age x Gender x Industry x County											
zero measured value, after rounding	0.2674	11,179,951	0	0.20	95.60%	(a)	(a)	(a)	9	1	(a)
1-2	0.0052	216,857	2	0.33	69.00%	0.1409	0.2958	0.7036	18	1	39.35%
3-9	0.2274	9,509,759	5	0.71	62.50%	0.0808	0.1639	0.3791	23	1	21.63%
10-99	0.3462	14,475,571	26	4.93	72.70%	0.0367	0.0805	0.1797	17	3	10.73%
100-999	0.1295	5,413,281	221	50.70	77.40%	0.0134	0.0295	0.0611	15	10	3.96%
+1000	0.0243	1,017,595	1,896	429.00	76.50%	0.0041	0.0099	0.0194	15	28	1.33%
Race x Ethnicity x Industry x County											
zero measured value, after rounding	0.6503	20,835,268	0	0.19	95.80%	(a)	(a)	(a)	9	1	(a)
1-2	0.0049	157,195	2	0.65	92.20%	0.2579	0.6021	0.8874	10	1	82.62%
3-9	0.1144	3,664,172	5	2.08	89.10%	0.1213	0.2990	0.5761	11	2	40.76%
10-99	0.1366	4,375,170	26	8.75	84.90%	0.0385	0.1074	0.2582	12	4	14.56%
100-999	0.0685	2,195,777	244	67.70	80.80%	0.0131	0.0306	0.0705	13	11	4.14%
+1000	0.0253	810,388	2,467	663.00	80.30%	0.0031	0.0093	0.0204	13	35	1.26%
Gender x Education x Industry x County											
zero measured value, after rounding	0.1317	2,774,036	0	0.26	95.30%	(a)	(a)	(a)	9	1	(a)
1-2	0.0059	124,663	2	1.32	93.80%	0.4278	0.6500	0.9368	10	2	89.19%
3-9	0.2237	4,709,531	5	3.83	93.50%	0.2365	0.3766	0.6065	10	3	51.68%
10-99	0.4122	8,679,900	27	20.60	93.60%	0.0860	0.1633	0.2954	10	6	22.40%
100-999	0.1839	3,871,290	230	181.00	93.80%	0.0288	0.0564	0.0958	10	18	7.74%
+1000	0.0426	897,250	2,011	1640.00	94.20%	0.0086	0.0192	0.0315	10	56	2.64%

Notes: Total employment is defined as all jobs held by a worker at the same establishment during the quarter. Statistics are computed across all state-year-quarters within a table. The "Private" category of establishments includes only private establishments. All tables include all valid QWI age groups with the exception of any table including education, in which case only jobs with workers age 25 and older are included. For statistic definitions for total employment, please see their respective equations in the accompanying text: Count 6, Total Variation 15, Missingness Ratio 16, Coefficient of Variation 32. (a) Undefined value.

Table A.9: Summary of Total Variability of Private Total Payroll (*Payroll*) by Table and Count

Table and *EmpTotal* count range	Proportion of Cells	Number of Cells	Median Payroll	Median Total Variation	Median Rubin Missingness Rate (Percent)	Quantiles of Coefficient of Variation			Median Approximate 90% Confidence Intervals Margin of Error		
						5th	Median	95th	Median df	Count	Percent
Private											
Age x Gender											
+1000	1.0000	46,480	375,627,224.50	3.96E+11	29.50%	0.0005	0.0016	0.0083	104	811,617.46	0.21%
Race x Ethnicity											
10-99	0.0199	695	233,792.00	2.29E+09	97.30%	0.1187	0.2083	0.4509	9	66,183.38	28.80%
100-999	0.1306	4,553	2,166,851.00	1.85E+10	96.10%	0.0334	0.0678	0.1493	9	188,112.25	9.38%
+1000	0.8495	29,612	71,561,892.50	4.67E+11	82.40%	0.0005	0.0070	0.0479	13	922,671.93	0.94%
Gender x Education											
+1000	1.0000	23,240	1,122,441,816.50	2.05E+13	96.10%	0.0018	0.0042	0.0117	9	6,261,928.94	0.57%
Industry x County											
zero measured value, after rounding	0.0040	12,955	0.00	9.45E+06	99.80%	0.0392	0.4429	1.0808	9	4,251.55	61.25%
1-2	0.1144	373,579	39,363.00	8.17E+06	0.00%	0.0000	0.0671	0.5900	9999	3,663.33	8.60%
3-9	0.0140	45,741	28,035.00	5.56E+06	0.00%	0.0302	0.0834	0.5458	9999	3,022.05	10.68%
10-99	0.2305	753,131	214,551.00	1.23E+08	9.57%	0.0198	0.0536	0.2327	984	14,222.64	6.88%
100-999	0.3899	1,273,670	1,568,970.00	2.55E+09	76.50%	0.0108	0.0307	0.0905	15	67,697.26	4.12%
+1000	0.2473	807,706	19,017,146.50	6.38E+10	80.70%	0.0035	0.0118	0.0329	13	341,035.20	1.59%
Age x Gender x Industry x County											
zero measured value, after rounding	0.1701	8,888,449	0.00	6.05E+05	100.00%	0.0000	0.2197	1.3084	9	1,075.74	30.39%
1-2	0.1618	8,454,917	4,523.00	3.34E+05	12.30%	0.0000	0.1107	0.8974	598	741.46	14.20%
3-9	0.1829	9,557,988	17,743.00	8.87E+06	83.20%	0.0453	0.1636	0.6023	13	4,021.15	22.08%
10-99	0.3198	16,713,425	113,931.00	1.16E+08	85.80%	0.0327	0.0939	0.2689	12	14,606.91	12.73%
100-999	0.1367	7,142,409	1,166,690.00	2.00E+09	87.20%	0.0144	0.0384	0.0942	11	60,974.46	5.23%
+1000	0.0289	1,511,324	14,164,728.00	3.60E+10	84.30%	0.0049	0.0135	0.0327	12	257,324.15	1.83%
Race x Ethnicity x Industry x County											
zero measured value, after rounding	0.4859	19,047,330	0.00	3.28E+06	100.00%	0.0115	0.4685	1.5608	9	2,504.77	64.80%
1-2	0.1727	6,771,506	4,934.00	6.96E+06	98.50%	0.0745	0.5842	1.2527	9	3,648.68	80.80%
3-9	0.1107	4,339,494	20,418.00	6.02E+07	96.70%	0.1040	0.4024	0.8577	9	10,730.73	55.65%
10-99	0.1337	5,240,825	125,020.00	3.38E+08	93.50%	0.0423	0.1488	0.3976	10	25,227.29	20.42%
100-999	0.0691	2,707,617	1,330,383.00	3.50E+09	88.60%	0.0153	0.0433	0.1145	11	80,661.63	5.90%
+1000	0.0279	1,094,612	16,504,365.00	5.75E+10	84.60%	0.0044	0.0135	0.0349	12	325,209.49	1.83%
Gender x Education x Industry x County											
zero measured value, after rounding	0.0845	2,207,640	153.00	2.35E+05	99.50%	0.0000	0.1886	2.1637	9	670.45	26.08%
1-2	0.1565	4,089,395	6,804.00	1.91E+07	98.80%	0.3127	0.6056	1.1470	9	6,044.33	83.76%
3-9	0.1716	4,484,091	25,232.00	1.38E+08	97.90%	0.2627	0.4790	0.8323	9	16,246.91	66.24%
10-99	0.3615	9,446,881	160,107.00	1.10E+09	97.20%	0.1026	0.2084	0.4189	9	45,869.87	28.83%
100-999	0.1797	4,695,684	1,522,171.00	1.37E+10	96.90%	0.0367	0.0763	0.1454	9	161,879.36	10.55%
+1000	0.0463	1,209,869	17,075,678.00	2.34E+11	96.40%	0.0118	0.0277	0.0559	9	669,020.05	3.83%

Notes: Total Payroll is defined only over total employment. It is calculated by summing the earnings for the reference quarter for total employment. See the table on total employment for the relevent counts. Statistics are computed across all state-year-quarters within a table. The "Private" category of establishments includes private only private establishments. All tables include all valid QWI age groups with the exception of any table including education, in which case only jobs with workers age 25 and older are included. For statistic definitions for beginning of quarter employment, please see their respective equations in the accompanying text: Total payroll 28, Total Variation 31, Missingness Ratio 16, Coefficient of Variation 32. (a) Undefined value.

Table A.10: Summary of Total Variability of Private Average Monthly Earnings (*EarnS*) by Table and Count

Table and *EmpS* count range	Proportion of Cells	Number of Cells	Median Average Monthly Earnings	Median Total Variation	Median Rubin Missingness Rate (Percent)	Quantiles of Coefficient of Variation			Median Approximate 90% Confidence Intervals Margin of Error		
						5th	Median	95th	Median df	Count	Percent
Private											
Age x Gender											
100-999	0.0001	6	1,686.00	13,600.00	87.00%	0.0691	0.0691	0.0691	11	159.00	9.42%
+1000	0.9999	44,938	2,146.00	6.79	22.90%	0.0004	0.0013	0.0066	171	3.35	0.17%
Race x Ethnicity											
3-9	0.0005	17	2,409.00	361,000.00	96.80%	0.1451	0.2600	0.6976	9	830.97	35.95%
10-99	0.0351	1,184	2,106.50	71,000.00	95.50%	0.0605	0.1252	0.3384	9	368.52	17.31%
100-999	0.1780	6,001	2,189.00	8,490.00	94.50%	0.0147	0.0425	0.1042	10	126.43	5.84%
+1000	0.7863	26,506	2,471.00	168.00	73.70%	0.0004	0.0052	0.0321	16	17.33	0.69%
Gender x Education											
+1000	1.0000	22,472	2,847.00	84.60	94.40%	0.0013	0.0032	0.0088	10	12.62	0.44%
Industry x County											
zero measured value, after rounding	0.0020	6,177	0.00	2,140,000.00	99.30%	(a)	(a)	(a)	9	2023.20	(a)
1-2	0.1096	342,768	2,088.00	7,230.00	0.00%	0.0000	0.0545	0.2903	9999	108.98	6.98%
3-9	0.0294	92,014	1,523.00	8,540.00	0.00%	0.0202	0.0655	0.2728	9999	118.44	8.39%
10-99	0.2836	886,839	1,957.00	5,240.00	11.20%	0.0150	0.0385	0.1256	714	92.85	4.93%
100-999	0.3749	1,172,234	2,265.00	2,020.00	65.80%	0.0081	0.0207	0.0551	20	59.57	2.75%
+1000	0.2005	626,901	2,701.00	414.00	70.60%	0.0026	0.0080	0.0216	18	27.07	1.07%
Age x Gender x Industry x County											
zero measured value, after rounding	0.0025	98,832	0.00	2,690,000.00	99.60%	(a)	(a)	(a)	9	2268.34	(a)
1-2	0.2269	8,953,462	1,297.00	9,140.00	0.00%	0.0000	0.0850	0.4718	9999	122.53	10.90%
3-9	0.2410	9,509,759	1,479.00	17,400.00	66.90%	0.0280	0.0932	0.2942	20	174.82	12.35%
10-99	0.3668	14,475,571	1,828.00	8,620.00	74.70%	0.0210	0.0538	0.1433	16	124.11	7.19%
100-999	0.1372	5,413,281	2,300.00	2,280.00	77.30%	0.0087	0.0224	0.0545	15	64.01	3.00%
+1000	0.0258	1,017,595	3,109.00	578.00	72.90%	0.0033	0.0084	0.0204	16	32.14	1.12%
Race x Ethnicity x Inudstry x County											
zero measured value, after rounding	0.0043	74,124	0.00	6,290,000.00	99.90%	(a)	(a)	(a)	9	3468.62	(a)
1-2	0.3501	5,991,260	1,835.00	226,000.00	97.20%	0.0499	0.2686	0.7331	9	657.48	37.15%
3-9	0.2141	3,664,172	1,942.00	126,000.00	93.50%	0.0563	0.1853	0.4744	10	487.08	25.42%
10-99	0.2557	4,375,170	2,082.00	24,500.00	86.80%	0.0246	0.0757	0.2127	11	213.41	10.33%
100-999	0.1283	2,195,777	2,290.00	3,170.00	79.90%	0.0097	0.0253	0.0661	14	75.73	3.40%
+1000	0.0474	810,388	2,723.00	508.00	75.50%	0.0032	0.0087	0.0221	15	30.22	1.16%
Gender x Education x Industry x County											
zero measured value, after rounding	0.0037	83,776	0.00	3,250,000.00	98.60%	(a)	(a)	(a)	9	2493.29	(a)
1-2	0.1917	4,326,849	1,800.00	424,000.00	98.10%	0.1419	0.3679	0.8320	9	900.56	50.88%
3-9	0.2087	4,709,531	1,908.00	241,000.00	95.90%	0.1254	0.2615	0.5459	9	678.95	36.16%
10-99	0.3846	8,679,900	2,210.00	63,700.00	94.40%	0.0534	0.1158	0.2588	10	346.32	15.89%
100-999	0.1715	3,871,290	2,558.00	12,000.00	94.30%	0.0203	0.0440	0.0937	10	150.32	6.04%
+1000	0.0398	897,250	3,175.00	2,870.00	94.20%	0.0076	0.0174	0.0408	10	73.51	2.38%

Notes: Average Monthly Earnings is defined only over full-quarter jobs. It is calculated by taking the earnings for the reference quarter for full-quarter jobs and dividing by 3. See the table on full-quarter employment for the relevent counts. Statistics are computed across all state-year-quarters within a table. The "Private" category of establishments includes private only private establishments.. All tables include all valid QWI age groups with the exception of any table including education, in which case only jobs with workers age 25 and older are included. For statistic definitions for beginning of quarter employment, please see their respective equations in the accompanying text: Average Monthly Earnings 20, Total Variation 23, Missingness Ratio 16, Coefficient of Variation 32. (a) Undefined value.

Table A.11: Between Variance of Beginning-of-Quarter (B) Population Counts

		Coefficient of Variation			
	Cell Count	Mean	Std Dev	Minimum	Maximum
A: Establishment Type and Age Range					
Population					
All Valid QWI Ages, All Establishments	2,957	1.059E-05	6.175E-06	1.738E-06	5.334E-05
All Valid QWI Ages, Private Establishments	2,957	1.187E-05	6.911E-06	1.984E-06	6.670E-05
B: State					
Postal Code					
AK	188	6.521E-05	5.021E-05	6.874E-06	2.234E-04
AL	172	3.040E-05	2.319E-05	3.695E-06	8.993E-05
AR	148	3.248E-05	2.201E-05	7.540E-06	7.745E-05
AZ	124	4.091E-05	3.357E-05	6.086E-06	1.385E-04
CA	324	2.463E-05	2.074E-05	2.581E-06	9.171E-05
CT	252	3.362E-05	2.643E-05	5.869E-06	1.417E-04
DC	104	8.532E-05	5.890E-05	1.611E-05	1.984E-04
DE	212	6.939E-05	5.802E-05	8.776E-06	2.319E-04
FL	304	1.812E-05	1.435E-05	2.625E-06	7.097E-05
GA	220	3.069E-05	2.497E-05	3.249E-06	1.027E-04
HI	256	4.458E-05	4.397E-05	5.854E-06	2.369E-04
IA	208	2.941E-05	2.218E-05	4.336E-06	1.009E-04
ID	332	6.875E-05	5.681E-05	6.696E-06	3.292E-04
IL	348	2.606E-05	2.101E-05	2.660E-06	7.849E-05
IN	220	2.383E-05	1.818E-05	2.140E-06	6.173E-05
KS	300	3.931E-05	3.122E-05	5.063E-06	1.262E-04
KY	172	2.659E-05	1.998E-05	3.718E-06	7.896E-05
LA	268	2.138E-05	1.482E-05	4.411E-06	6.201E-05
MD	348	3.166E-05	2.787E-05	3.609E-06	1.276E-04
ME	248	3.069E-05	2.227E-05	5.408E-06	9.439E-05
MI	180	1.896E-05	1.498E-05	3.122E-06	5.559E-05
MN	276	2.373E-05	1.933E-05	3.230E-06	6.993E-05
MO	268	2.291E-05	1.892E-05	2.885E-06	6.340E-05
MS	132	3.457E-05	2.448E-05	5.182E-06	8.206E-05
MT	300	4.375E-05	3.270E-05	7.252E-06	1.399E-04
ND	220	4.782E-05	3.732E-05	6.366E-06	1.632E-04
NE	204	3.910E-05	3.037E-05	6.833E-06	1.090E-04
NH	140	4.042E-05	2.809E-05	7.394E-06	9.946E-05
NJ	252	2.982E-05	2.341E-05	3.494E-06	1.275E-04
NM	260	7.141E-05	6.564E-05	6.204E-06	3.728E-04
NV	220	6.333E-05	5.146E-05	6.337E-06	1.739E-04
NY	188	2.334E-05	2.095E-05	2.948E-06	1.143E-04
OH	188	1.475E-05	1.147E-05	2.241E-06	4.309E-05
OK	188	4.435E-05	3.440E-05	4.895E-06	1.131E-04
OR	332	3.848E-05	2.973E-05	5.838E-06	1.269E-04
PA	236	1.181E-05	8.594E-06	1.738E-06	3.660E-05
RI	268	6.231E-05	4.507E-05	6.479E-06	1.753E-04
SC	220	3.604E-05	2.772E-05	4.599E-06	9.803E-05
SD	220	5.157E-05	4.009E-05	6.421E-06	1.497E-04
TN	220	2.394E-05	1.913E-05	2.935E-06	8.213E-05
TX	268	1.945E-05	1.583E-05	2.284E-06	7.985E-05
UT	196	6.618E-05	5.351E-05	7.202E-06	1.792E-04
VA	220	2.814E-05	2.341E-05	4.003E-06	1.111E-04
VT	188	4.439E-05	3.280E-05	5.941E-06	1.317E-04
WA	348	3.202E-05	2.560E-05	4.074E-06	9.715E-05
WI	348	2.128E-05	1.751E-05	1.960E-06	7.021E-05
WV	236	2.298E-05	1.554E-05	3.872E-06	7.009E-05
WY	172	8.874E-05	6.879E-05	1.586E-05	2.701E-04

Notes: There is small amount of between-implicate variance of state counts for beginning-of-quarter employment. We summarize the between variance using the coefficient of variation defined as the square root of the between-implicate variance divided by the average between-implicate weighted counts. Panel A summarizes the coefficient of variation for the between variance for the four different types of ownership type and age populations. The summary is taken across all state-year-quarters. Panel B summarizes the coefficient of variation for all states across all year, quarters, and ownership types and age range combinations.

Table A.12: Between Variance of Full-Quarter (F) Population Counts

		Coefficient of Variation			
	Cell Count	Mean	Std Dev	Minimum	Maximum
A: Establishment Type and Age Range					
Population					
All Valid QWI Ages, All Establishments	2,957	1.027E-05	5.891E-06	2.149E-06	5.356E-05
All Valid QWI Ages, Private Establishments	2,957	1.152E-05	6.592E-06	1.990E-06	5.403E-05
B: State					
Postal Code					
AK	188	5.625E-05	4.128E-05	7.485E-06	1.601E-04
AL	172	2.873E-05	2.214E-05	3.299E-06	8.659E-05
AR	148	3.144E-05	2.094E-05	6.305E-06	7.746E-05
AZ	124	4.139E-05	3.343E-05	5.211E-06	1.432E-04
CA	324	2.273E-05	1.914E-05	2.422E-06	8.591E-05
CT	252	3.258E-05	2.605E-05	5.610E-06	1.371E-04
DC	104	8.303E-05	5.866E-05	1.708E-05	2.181E-04
DE	212	6.386E-05	5.424E-05	5.393E-06	2.005E-04
FL	304	1.645E-05	1.296E-05	2.559E-06	6.220E-05
GA	220	2.916E-05	2.281E-05	3.254E-06	8.538E-05
HI	256	4.090E-05	4.112E-05	5.684E-06	2.272E-04
IA	208	2.801E-05	2.095E-05	4.748E-06	7.924E-05
ID	332	6.090E-05	4.640E-05	8.729E-06	1.853E-04
IL	348	2.404E-05	1.901E-05	2.408E-06	6.948E-05
IN	220	2.254E-05	1.722E-05	3.352E-06	6.253E-05
KS	300	3.776E-05	3.054E-05	5.557E-06	1.261E-04
KY	172	2.588E-05	1.884E-05	3.681E-06	8.234E-05
LA	268	2.087E-05	1.429E-05	3.630E-06	5.697E-05
MD	348	2.926E-05	2.510E-05	3.640E-06	1.148E-04
ME	248	2.783E-05	1.945E-05	4.771E-06	8.563E-05
MI	180	1.626E-05	1.206E-05	2.406E-06	4.997E-05
MN	276	2.264E-05	1.861E-05	2.661E-06	6.325E-05
MO	268	2.164E-05	1.737E-05	3.001E-06	6.110E-05
MS	132	3.374E-05	2.393E-05	4.847E-06	9.564E-05
MT	300	4.097E-05	2.925E-05	7.754E-06	1.329E-04
ND	220	4.407E-05	3.356E-05	5.709E-06	1.191E-04
NE	204	3.838E-05	2.992E-05	4.032E-06	1.109E-04
NH	140	3.957E-05	2.731E-05	6.623E-06	9.685E-05
NJ	252	2.814E-05	2.202E-05	4.148E-06	9.915E-05
NM	260	6.823E-05	5.648E-05	6.973E-06	2.258E-04
NV	220	5.935E-05	4.752E-05	6.652E-06	1.766E-04
NY	188	2.177E-05	1.894E-05	2.341E-06	9.451E-05
OH	188	1.402E-05	1.088E-05	2.160E-06	3.793E-05
OK	188	4.342E-05	3.421E-05	3.832E-06	1.146E-04
OR	332	3.542E-05	2.710E-05	5.739E-06	1.056E-04
PA	236	1.094E-05	7.845E-06	1.990E-06	3.553E-05
RI	268	5.713E-05	4.069E-05	9.523E-06	1.553E-04
SC	220	3.390E-05	2.587E-05	4.492E-06	1.016E-04
SD	220	4.917E-05	3.734E-05	7.047E-06	1.459E-04
TN	220	2.221E-05	1.741E-05	2.674E-06	6.856E-05
TX	268	1.757E-05	1.413E-05	2.003E-06	6.805E-05
UT	196	6.683E-05	5.510E-05	7.998E-06	1.977E-04
VA	220	2.636E-05	2.214E-05	3.717E-06	1.089E-04
VT	188	4.051E-05	2.909E-05	7.829E-06	1.249E-04
WA	348	2.724E-05	2.122E-05	3.714E-06	7.521E-05
WI	348	2.052E-05	1.673E-05	3.026E-06	6.773E-05
WV	236	2.199E-05	1.394E-05	3.598E-06	6.286E-05
WY	172	8.041E-05	6.154E-05	1.300E-05	2.730E-04

Notes: There is small amount of between-implicate variance of state counts for full-quarter employment. We summarize the between variance using the coefficient of variation defined as the square root of the between-implicate variance divided by the average between-implicate weighted counts. Panel A summarizes the coefficient of variation for the between variance for the four different types of ownership type and age populations. The summary is taken across all state-year-quarters. Panel B summarizes the coefficient of variation for all states across all year, quarters, and ownership types and age range combinations.