Westat®

Improving lives through research™

Final Methodology Report for the OSHA/SOII Initiative on Combining Survey and Administrative Records

Task Order 2, BLS BPA 1625DC-17-A-0001

Authors

Lou Rizzo
J. Michael Brick

April 3, 2018

Prepared for:
Bureau of Labor Statistics
2 Massachusetts Ave. NE
Washington, DC 20212

Prepared by:
Westat
An Employee-Owned Research Corporation®
1600 Research Boulevard
Rockville, Maryland 20850-3129
(301) 251-1500

Table of Contents

Appendixes **Page**

1 General Introduction

This methodology report presents "a way forward" for utilizing the new OSHA electronic data collection initiative (called the 'OSHA EDCI', with 'OSHA EDC' for OSHA electronic data collection) in the context of SOII. This way forward sets out plausible scenarios for the rollout of this initiative, discusses options for SOII data collection and estimation methods under these scenarios, and discusses aspects that should be evaluated before making changes to the SOII as this initiative evolves.

Section 2 presents an overview of SOII, presenting aspects that are directly relevant to this methodology report, and gives possible scenarios for the progress of this initiative. The scenario we anticipate is that in the initial stages the assumption will be made that there is no relative bias in the OSHA EDCI estimates, which will allow for a composite estimation approach. At the same time, an empirical evaluation of the validity of this assumption is important.

Section 3 presents the composite estimation approach under the assumption that there is no relative bias in the OSHA EDCI estimates. Section 4 presents a methodology for checking this relative bias assumption through a variety of area-level and unit-level analyses. Section 5 provides an alternative estimation approach allowing for relative bias in the OSHA EDCI estimates. Section 6 discusses these methodologies.

Appendix A provides details of the composite estimation approach for quantiles, and Appendix B gives details for the linkage methods.

Overview of the OSHA Initiative and of the SOII Sample Design, Questionnaire, and Estimation 2

In Sections 2.1, 2.2, and 2.3, we review aspects of the SOII sample design and the OSHA EDCI that are relevant for this methodology report. In Section 2.4, we put forward a series of scenarios which represent our best guess as to how this initiative may roll out (in branched scenarios).

2.1 SOII Sample Design, Questionnaire, and Estimation

This sample design overview of SOII is based on Selby et al. (2008). The universe for the SOII sample consists of all private sector employers and state and local governments, excluding self-employed persons and small farms, as well as mining and railroad industries. The frame for the sample of establishments is the BLS Quarterly Census of Employment and Wages (QCEW), although some of the states use their own frames of state and local government agencies.

The sample design is a stratified simple random sample, with stratification defined by state, ownership (state government, local government, private industry), NAICS industry (divided into strata called Target Estimation Industries, or TEIs), and employment size class. The sample sizes n_h for each stratum cell are proportional within each state to

$$n_h \propto E_h\sqrt{p_h(1-p_h)}$$

where E_h is total employment in stratum cell h, and p_h is the TRC rate (the ratio per establishment of total recordable cases[1] for a year to total annual employee hours) for the stratum. It should be noted that this sample design will oversample larger firms (those with greater employment) and also, to a somewhat lesser degree, firms with a high incidence of total recordable cases. But every establishment with at least one employee, whether or not there were occupational injury or illness incidents (OII), will have a chance of selection.

[1] Total recordable cases are total cases with day away from work (DAFW) plus cases with days of job transfer or restriction (DJTR), plus other recordable cases. A recordable case is generally a case in which there were DAFW, DJTR, injuries and illnesses involving medical treatment beyond first aid, and also some other related scenarios. ~~in which a case becomes recordable even with no DAFW or DJTR (for example, if there was a trip to the hospital emergency department)~~.

Once an establishment is sampled, they are asked to provide all data on non-fatal injuries and illnesses for a full calendar year (January to December). There is subsampling of employees or time periods for units reporting more than 15 Days Away From Work (DAFW) cases, and no subsampling for all other units (i.e., a census is taken). The data collection covers the full year's incidences of nonfatal occupational injuries or illnesses within the establishment.

The SOII questionnaire mail-out includes the name and address of the establishment and the establishment ID (EIN). The questionnaire form itself asks the sampled establishment about these items at the establishment level:

- Annual average number of employees and total hours worked;
- Special conditions affecting employment (e.g., strikes);
- Total number of nonfatal OII cases with at least one day away from work, total number of OII cases with job transfers or restrictions, and total number of other nonfatal OII recordable cases.
- Total number of days away from work and total number of days with job transfer or restriction; and
- Injury and illness types.

Each recordable case with days away from work (DAFW) has information recorded about it[2] including:

- Date of OII;
- Number of days (away from work, or with job transfer or restriction);
- Whether employee was treated in emergency room, and/or hospitalized overnight;
- Details (with open-ended questions with text answers) about what the employee was doing right before the incident, what happened, what was the OII, and what object harmed the employee.

The same information is collected in SOII for recordable cases with days with job transfers or restrictions (DJTR) for establishments for selected industries in the private sector.

[2] It should be noted that establishments with up to 15 DAFW cases (in a year) provide detail about all of these cases, but establishments with more than 15 DAFW cases provide detail only on a subsample of their cases.

Also included on the questionnaire are information about the employee suffering the OII including:

- Name of employee;
- Job title;
- Type of work;
- Employee's race and ethnic background;
- Employee's age; and
- Employee's gender.

SOII estimates include both estimates based on Form 300A establishment counts, and estimates based on Form 300/301 case-level data. Among the estimates that SOII has produced and wishes to produce in the future are the following:

- Injury rates by domain;
- Case totals by domain; and
- Establishment injury rate quartiles within a domain.

Domains are defined by private industry/state or local government, at the state level, at the industry level (NAICS categories), and by establishment size within industry level. Domain-level injury rates are defined to be total cases in the domain divided by total hours worked in the domain. The totals should be consistent (additive) when domains are combined into higher-level domains. This will occur automatically, for example, if the totals are generated by aggregating establishment counts using a single set of unit weights. These estimates should be approximately unbiased at each relevant domain level.

2.2 Overview of the OSHA EDC Initiative

The OSHA EDCI was originally planned to ask all establishments with 250 or more employees to provide their OSHA information electronically for publication on a public-use OSHA website. Establishments with 20 to 249 employees that are in industries with historically high rates of OII are also required to submit their OSHA information electronically[3].

At least in the initial phases of the OSHA EDC initiative, Form 300-301 case-level information is not planned to be collected electronically for posting on the website. This may change in later phases of the initiative. In all phases of the initiative, establishments with 20 to 249 employees in historically high-risk industries will be required to submit their Form 300A aggregate totals electronically.

The OSHA Form 300A which will be collected electronically provides the following information at the establishment level:

- NAICS classification
- Total employees and total hours[4]
- Total number of DAFW, DJTR, other cases
- Total number of OII by injury and illness type

[3] When specifying establishment size for deciding on eligibility, the employment count considered is the peak employment count for the establishment during the year rather than the average annual employment. This means some firms below the cutoff (20 or 250 employees) as per their average employment may in fact be eligible as their peak employment exceeds the cutoff. This will tend to increase marginally the set of establishments eligible for OSHA data collection.

[4] Note that total employees collected on OSHA Form 300A is an average number of employees as in SOII.

At a later point, establishments with 250 or more employees may be required to also submit electronically information from OSHA Forms 300 and 301. OSHA Form 300 includes the following information about each recordable case:

- Name of employee;
- Job title;
- Date of OII;
- An open-ended description (with text answers) of what the employee was doing, what happened, parts of body affected, object or substance injuring the employee;
- DAFW, DJTR, or other type of case;
- Number of days DAFW or DJTR; and
- Type of OII (injury, skin disorder, etc.).

OSHA Form 301 provides the following information about each OII individually:

- Information about employee including name, address, date of birth, date hired, and gender[5];
- Information about health care professional providing care for OII including name, facility, whether or not emergency room care was provided, or overnight hospitalization; and
- Detailed information (using open-ended questions with text answers) about the OII including date and time of injury, and text information about what the employee was doing, information about the OII, the nature of the injury or illness, and object or substance harming the employee.

[5] If OSHA case records are electronically reported in the future, it is unlikely that names, addresses, and other information that could reveal identity of individuals would be reported on the public-use website, to protect confidentiality of individual persons.

2.3 Further Aspects Regarding the SOII and OSHA EDCI

The SOII universe is more complete, including all establishments with at least one employee, and including state and local government entities, but it is a sample. The sampling process includes a higher proportion of larger establishments and establishments in industries with a historically higher incidence of OII, but all establishments have some chance of selection.

The OSHA EDCI universe on the other hand is restricted. It only includes private establishments. Establishments with fewer than 25 employees are excluded entirely. Establishments with 25 to 249 employees who are not in historically high-OII incidence industries are also excluded, and establishments with 25 to 249 employees who are in historically high-OII incidence industries are only required to submit the summary data (OSHA Form 300A). Within the targeted universe, the canvassing is a census.

Table 1 derived from Pierce (2017) provides an analysis of the potential overlap between a SOII sample and the OSHA EDCI universe, assuming the OSHA initiative is carried out as planned and the SOII sample design remains unchanged. About 40% of SOII sample units will overlap with OSHA EDCI establishments, representing a weighted percentage of 8%.

Table 1. Expected percentages of a future SOII sample with OSHA EDCI establishments

	Establishments	Reported employment	Total recordable cases
Unweighted percentage distribution			
No OSHA electronic reporting	**60.4**	**30.2**	**13.3**
OSHA electronic reporting	**39.6**	**69.8**	**86.7**
Weighted distribution			
No OSHA electronic reporting	**92.3**	**58.2**	**35.0**
OSHA electronic reporting	**7.7**	**41.8**	**65.0**

The SOII questionnaire collects a set of information very similar to that collected in OSHA Forms 300, 300A, and 301. The SOII Injury and Illness Case Form essentially collects the same information as OSHA Form 300 and 301, with a few differences. The SOII instrument collects information about race/ethnicity that is not collected in the OSHA instruments. The OSHA instrument collects information about the physician providing treatment. SOII Sections 1 and 2 collect establishment and OII summary information very similar to OSHA Form 300A. The SOII instrument collects extra information about special conditions affecting the establishment in that year (strikes, natural

disasters, economically driven work speedups or slowdowns), but this is only used in analysis and not in estimation.

2.4 Potential Scenarios for the Rollout of the OSHA EDC Initiative

In this section, we speculate on a number of possible scenarios for how the OSHA EDCI and SOII may develop over time. A basic branch is between 'concordance' – a documented lack of relative measurement error bias in the OSHA EDCI estimates as compared to SOII estimates, and 'discordance' – evidence of possible measurement error bias.

2.4.1 Year 1 of the OSHA EDCI Rollout

In Year 1, OSHA EDCI will be initiated and we would assume that SOII will continue with its sample design and data collection as is. We assume that OSHA EDCI in this first year will only collect Form 300A type totals information. It may be that establishments will notice the duplicated data collection and question it, which may disturb response rates to SOII collection. This should be monitored. If EIN is available in the OSHA EDCI, then linking between the data collections corresponding to the same establishments may be straightforward. If not, then a procedure as described in Section 4 and Appendix B could be developed to provide linking that can generate a paired data set that can be analyzed.

Response rates to OSHA EDCI will be monitored throughout this first year. If the response rates to this initiative are sufficiently high then we assume the initiative can continue as is to a second year (see Section 2.4.2 below). If the response rates fall below a reasonable threshold, then the initiative will need to re-evaluate its procedures and re-start the initiative to 'Year 1' again. Progress to 'Year 2' rather than a second try at 'Year 1' occurs if the OSHA EDCI response rates are judged to be at an acceptable level (a judgment made by OSHA and BLS management). While no definite rules exist for establishing adequate response rates to minimize nonresponse bias, one option that could be considered is examining the differential in the response rates between the two surveys. If the OSHA EDCI rates by stratum are similar to those of SOII, then they might be considered adequate. If the rates differ substantially (say 5 percentage points or more) then some corrective actions to increase the rates might be deemed necessary.

2.4.2 Year 2 of the OSHA EDCI Rollout

In Year 2, a second year of OSHA EDCI data collection will be done with SOII also continuing forward with its current design and procedures. This second year will also include a processing of the OSHA EDCI outcome data collected from Year 1. Nonresponse adjustments should be computed for nonresponse to OSHA EDCI following a nonresponse analysis. This will lead to OSHA EDCI estimates for Form 300A items for the 'OSHA EDCI domains': the industry and size domains covered by the OSHA EDC Initiative.

At this point, a thorough bias analysis should be done to check for relative measurement error bias in the OSHA EDCI data. This can be at the macro level (comparing the domain-level Form 300A OSHA EDCI estimates with the corresponding SOII Form 300A estimates within the same domain: as discussed in Section 4.2), and/or at the micro level (comparing Form 300A OSHA EDCI records with corresponding linked SOII Form 300A records, as covered in Section 4.3).

If the bias analysis strongly shows a universal lack of relative measurement error bias in the OSHA EDCI data, then we suggest moving forward in the 'concordance scenario'. A computational machinery can be developed for composite estimators and calibration to combine the OSHA EDCI and SOII estimates because they appear to have the same expectation. Form 300A estimates in overlapping OSHA-SOII domains should be composited. Form 300/301 SOII estimates in overlapping domains should be calibrated to OSHA Form 300A estimates. It may be that composite estimators can be generated for Year 1 based on composite estimation as early as Year 3.

If the bias analysis is more ambiguous, or shows the presence of relative measurement error bias in the OSHA EDCI data, then the road forward is much more difficult. Publication of simple composite estimators cannot be justified. One path would be to reconfigure the OSHA EDC to reduce or eliminate the measurement error bias. Options that OSHA might consider include a major revision of the data collection for the OSHA EDCI, or possibly including an enforcement mechanism for compliance. Thus in a sense we return to 'Year 1' as OSHA EDCI is reset. A second path (with some serious concerns from our perspective) is to accept the possibility of relative measurement error bias and to proceed with estimation schemes which allow for relative bias (see Section 5). A concern with this approach is that the OSHA website would still show the reported data that appear to have substantial measurement error and could lead to analysts producing biased

estimates using these data. Section 5 methods will not fully resolve the bias and are also complicated to implement.

2.4.3 Year 3 of the OSHA EDCI Rollout: Concordance Scenario

Under the concordance scenario, Year 3 can see publication of composite estimates for Form 300A data in OSHA-SOII overlapping domains from Year 1. Data collection in both OSHA EDCI and SOII will continue. Though Year 2 showed no measurement error bias, the bias analyses done in Year 2 to establish this (see Sections 4.2 and 4.3) should be continued to affirm a lack of bias. Calibration of SOII Form 300/301 estimates to OSHA EDCI Form 300A estimates (or technically the Form 300A composites) should be done as well.

2.4.4 Year 3 of the OSHA EDCI Rollout: Discordance Scenario

Under the discordance scenario and assuming that the biases in the OSHA EDCI are 'accepted', Empirical Bayes type estimation schemes could be developed to combine the OSHA and the SOII Form 300A estimates in overlapping domains, following the methodology described in Section 5. There would be no calibration of SOII Form 300/301 estimates to OSHA in this instance. Bias analysis would need to continue to allow for generation of bias models that can inform Empirical Bayes.

2.4.5 Years 4 and Beyond of the OSHA EDCI Rollout: Concordance Scenario

Years 4 and beyond under the concordance scenario will continue with composite estimation for Form 300A data, where there is overlapping data collection. Calibration of SOII Form 300/301 data to OSHA-SOII Form 300A will continue in domains where OSHA data collection takes place, as SOII will continue Form 300/301 data collection. Bias analysis to confirm the lack of relative measurement error bias should continue, though micro-level analysis is not feasible as SOII discontinues redundant data collection. Some comparisons at the macro-level (Section 4.2) might still be possible. At some point, it may no longer be necessary to continue to check for relative measurement error bias.

A final potential possibility is that OSHA EDCI could include Form 300/301 data in addition to Form 300A data. If this happens, a new process of checking for bias and composite estimation at the Form 300/301 level will be needed. Assuming a lack of relative measurement error bias is confirmed here as well, composite estimation can move forward. Calibration of Form 300/301 SOII to Form 300A OSHA could be discontinued. Finally, SOII may be completely discontinued in domains covered by OSHA EDCI Form 300/301 data collection. OSHA EDCI becomes a certainty stratum in SOII in a sense.

Composite Estimation and Calibration Assuming Unbiasedness in OSHA EDCI 3

This section describes composite estimation for combining OSHA EDCI and SOII estimates in circumstances in which both the domains and the questionnaires overlap. It is assumed that both the OSHA EDCI and SOII estimates have been adjusted through weighting adjustments for OSHA EDCI and SOII nonresponse, respectively. It is also assumed that measurement error bias has been checked, and OSHA EDCI has been confirmed not to suffer from relative measurement bias as compared to SOII. Section 3.1 describes composite estimation and Section 3.2 describes calibration. Appendix A covers the special case of quantile estimation.

3.1. Composite Estimators

Write $t = 1,..,T$ as the state, $i = 1,..,\ I_t$ as the NAICS industry domain within state (which may differ in definition across the states, so that I_t may not be the same in each state t). Write $k = 1,..,K$ as ownership and size-class stratum:

Ownership/Size Class Stratum	Ownership	Size class
k=1	Private	1 to 19 employees
k=2	Private	20 to 249 employees
k=3	Private	250 or more employees
k=4	State and Local Govt	All sizes

Write the domain population total of interest for state t, industry domain i:

$$Y_{ti} = \sum_{k=1}^{4} Y_{tik}$$

Note that Y_{tik} may be zero for many combinations of industry and ownership/size class. SOII will produce estimates for all non-zero Y_{tik} ($\hat{Y}_{tik}^{SOII}$). Assume $\hat{Y}_{tik}^{SOII}$ contains all of the weighting adjustments for SOII and can be considered an unbiased estimate of Y_{tik}.

OSHA data can be used to estimate $\hat{Y}_{tik}^{OSHA}$, but only for particular industries and size classes. Write $I(H)$ as high-risk industries for which smaller establishments (20 to 249 employees) are required to produce summary-level electronic information. $I(L)$ is the complement set.

Write $\hat{Y}_{tik}$ as the composite estimator for state t, industry domain i, ownership/size class k[6]. For $k = 1$, $k = 4$, and $k = 2$ within $I(L)$, only the SOII estimate can be computed, so that the composite estimator $\hat{Y}_{tik}$ is equal to $\hat{Y}_{tik}^{SOII}$. For $k = 3$, and $k = 2$ within $I(H)$, we have both $\hat{Y}_{tik}^{SOII}$ and a $\hat{Y}_{tik}^{OSHA}$ that can be composited. $\hat{Y}_{tik}^{OSHA}$ should be weight-adjusted for nonresponse and, if possible, calibrated as well to auxiliary control totals from the SOII sampling frame. There are no base weights in this case, as OSHA electronic forms are required of all eligible establishments.

We have then

$$\hat{Y}_{ti} = \sum_{k=1}^{4} \hat{Y}_{tik}$$

If $\hat{Y}_{tik}^{OSHA}$ is unbiased, (or unbiased after sufficient calibration has been done within the stratum tik), then simple composite estimators as in Lohr (2011) will be appropriate. For $i \in I(L)$ we have:

$$\hat{Y}_{ti} = \sum_{k\neq 3} \hat{Y}_{tik}^{SOII} + \left\{ w_{ti3}^{SOII} \hat{Y}_{ti3}^{SOII} + \left(1 - w_{ti3}^{SOII}\right) \hat{Y}_{ti3}^{OSHA} \right\}$$

where w_{ti3}^{SOII} and $1 - w_{ti3}^{SOII}$ $\left(0 \leq w_{ti3}^{SOII} \leq 1\right)$ are composite weights assigned to $\hat{Y}_{ti3}^{SOII}$ and $\hat{Y}_{ti3}^{OSHA}$, respectively, based on relative precision of the two estimates. This follows standard composite estimation methodology when both estimates are unbiased. For $i \in I(H)$ we have:

$$\hat{Y}_{ti} = \sum_{k=1,4} \hat{Y}_{tik}^{SOII} + \sum_{k=2,3} \left\{ w_{tik}^{SOII} \hat{Y}_{tik}^{SOII} + \left(1 - w_{tik}^{SOII}\right) \hat{Y}_{tik}^{OSHA} \right\}$$

The relative weights should be based on sample sizes and estimated design effects. In cases where we have both estimators we can write the variances of $\hat{Y}_{tik}^{SOII}$ and $\hat{Y}_{tik}^{OSHA}$ as follows:

[6] The size class strata for SOII sampling is (from Selby et al. (2008)) are 10 or less employees, 11 to 49 employees, 50 to 249 employees, 250 to 999 employees, and 1,000 or more employees. These sampling strata do not completely align with the strata arising from the OSHA initiative. A final stratification can be created in production which consists of the intersections: 10 or less employees, 11 to 19 employees, 20 to 49 employees, 50 to 249 employees, 250 to 999 employees, 1,000 or more employees.

$$Var\left(\hat{Y}_{tik}^{SOII}\right) = \frac{\sigma_y^2(tik)d_{tik}^{SOII}}{n_{tik}^{SOII}} \qquad Var\left(\hat{Y}_{tik}^{OSHA}\right) = \frac{\sigma_y^2(tik)d_{tik}^{OSHA}}{n_{tik}^{OSHA}}$$

The parameter $\sigma_y^2(tik)$ is a population-level variance for Y which is the same for the SOII estimate and the OSHA EDCI estimate (depending on the population within the particular domain tik). When computing the composite weights, w_{tik}^{SOII}, this variance parameter's value is ultimately irrelevant as it drops out when the composite weights are calculated (see below). n_{tik}^{SOII} and n_{tik}^{OSHA} are sample sizes for SOII and OSHA EDCI, respectively within domain tik, and d_{tik}^{SOII} and d_{tik}^{OSHA} are design effects for SOII and OSHA EDCI. These design effects should include any finite population corrections if these are nonnegligible (for stratified simple random samples these corrections are $1 - n_h/N_h$ where the subscripts indicate stratum *h*)). These are the ratios of the variances for SOII and OSHA EDCI respectively to the variance under a simple random sample of size n_{tik}^{SOII} (n_{tik}^{OSHA}) from the same population. See for example Valliant et al. (2013), Section 3.6.

Following Lohr (2011), the optimal weights are based on ratios of variances:

$$w_{tik}^{SOII} = \frac{Var\left(\hat{Y}_{tik}^{OSHA}\right)}{Var\left(\hat{Y}_{tik}^{SOII}\right)+Var\left(\hat{Y}_{tik}^{OSHA}\right)}.$$

Seeing that the $\sigma_y^2(tik)$ drops out of the formula we have:

$$w_{tik}^{SOII} = \frac{\frac{d_{tik}^{OSHA}}{n_{tik}^{OSHA}}}{\frac{d_{tik}^{SOII}}{n_{tik}^{SOII}}+\frac{d_{tik}^{OSHA}}{n_{tik}^{OSHA}}}.$$

Given the simplicity of the SOII sample design, the design effects d_{tik}^{SOII} will be driven primarily by weighting effects: the effects on variance of the SOII interviews from differential base weights and differential nonresponse adjustments. A widely-used formula for the design effect induced from weights assigned to adjust for sampling and nonresponse is the Kish factor

$$Deff_w = 1 + CV^2(w)$$

where $CV^2(w)$ is the square of the coefficient of variation of the weights (the standard deviation of the weights divided by the mean weight). See Kish (1992). This implicitly assumes a model where the weights are unrelated to the Y values. An appropriate value for the design effect should be based on an empirical evaluation of estimated variances $v\left(\hat{Y}_{tik}^{SOII}\right)$. The general hope is that a single value d_{tik}^{SOII} will be appropriate for the range of possible Y variables so that different composite weights

do not need to be used for different Y values. This is a topic for empirical research. As long as both the SOII and OSHA estimates are unbiased, any error in the w_{tik}^{SOII} values do not result in bias: it is just that the lowest possible variance may not be achieved.

For OSHA EDCI d_{tik}^{OSHA}, there is no sampling and no base weights, but there will be nonresponse adjustments assigned to adjust for noncooperating units. Computing variances for the OSHA EDCI estimates may be somewhat awkward as there is no sampling—only nonresponse—but a pseudo-randomization variance could be computed so that the uncertainty in the OSHA EDCI estimator can be evaluated. The pseudo-randomization variance is based on the underlying model that response is a simple random sample with equal probability within response cells with uniform response propensity (see for example Oh and Scheuren (1983)).

3.2. SOII Estimators for Case-Level Estimates

As noted above, the plan is that no electronic data will be collected for Form 300 or 301 type individual employee case-level data for any establishments, at least in the initial development of the OSHA EDCI. For any estimates that are based on this case-level data (such as estimates for subgroups of employees and estimates for types of treatment such as hospitalization), there will be no OSHA EDCI data, and the SOII alone will provide the estimates. The case-level estimates for each state and industry domain will be

$$\hat{Z}_{ti}^{SOII-ONLY} = \sum_{k=1}^{4} \hat{Z}_{tik}^{SOII}$$

Write $\hat{\mathbf{Z}}^{\mathbf{SOII-ONLY}}$ as a vector of estimates $\hat{Z}_{ti}^{SOII-ONLY}$ over all states $t = 1, \ldots, T$ and all industries $i = 1, \ldots I$. Write $\hat{\mathbf{X}}^{\mathbf{SOII-OSHA}}$ as composite estimates for Form 300A type OII totals derived from both SOII and OSHA sources using the composite estimators specified in Section 3.1. Write $\hat{\mathbf{X}}^{\mathbf{SOII-ONLY}}$ as estimates for these same OII totals based only on SOII, paralleling the $\hat{\mathbf{Z}}^{\mathbf{SOII-ONLY}}$ estimates for the case-level Form 300-301 information. The vector components for X are chosen on the basis of being correlated well with $\hat{\mathbf{Z}}^{\mathbf{SOII-ONLY}}$. The final calibrated version can then be written in GREG form as

$$\hat{\mathbf{Z}}^{\mathbf{FINAL}} = \hat{\mathbf{Z}}^{\mathbf{SOII-ONLY}} + \left(\hat{\mathbf{X}}^{\mathbf{SOII-OSHA}} - \hat{\mathbf{X}}^{\mathbf{SOII-ONLY}}\right)\hat{\boldsymbol{\beta}}$$

with $\widehat{\boldsymbol{\beta}}$ a parameter vector estimate regressing Z on X (see for example Valliant et al. (2014) Section 4.3). Note that the actual procedure would involve adjustments to the weights through poststratification or raking to control totals based on $\widehat{\mathbf{X}}^{\mathbf{SOII-OSHA}}$ as is standard for calibration.

This methodology is appropriate only when the OSHA EDCI Form 300A estimates are unbiased, so that the composite estimators $\widehat{\mathbf{X}}^{\mathbf{SOII-OSHA}}$ are also unbiased. Using biased control totals would be bias-inducing and counterproductive. Alternatively, if the OSHA EDCI estimates are biased, but a strategy such as that given in Section 5 can 'restore' unbiasedness, then it might be considered to also use the resultant $\widehat{\mathbf{X}}^{\mathbf{SOII-OSHA}}$ as control totals.

If the $\widehat{\mathbf{X}}^{\mathbf{SOII-OSHA}}$ are used as control totals, then variance estimation should take into account the variability of these control totals as they are estimates themselves. For example, one of the approaches in Dever and Valliant (2010) can be considered.

Checking For Relative Bias in OSHA EDCI 4

Confidence in the unbiasedness of the OSHA EDCI estimates will allow for complete inclusion of these estimates into the SOII estimation system using fairly straightforward methodologies, considerably improving SOII estimates. On the other hand, relative bias of these estimates will make it difficult to incorporate them without the use of complex methods that may ultimately fail to adjust out the biases. An evaluation of the potential bias is a critical task for this initiative.

Section 4.1 provides a theoretical analysis of the important distinction between nonresponse bias (bias from a mis-alignment of the sample of establishments with the population of establishments), and measurement error bias (a misrecording of data for a particular establishment). Section 4.2 describes 'macro-analysis' of measurement error bias, and Section 4.3 describes 'micro-analysis' of measurement error bias.

4.1 Nonresponse Bias vs. Measurement Error Bias

The critical issue is differences in relative biases in the two data sources SOII and OSHA EDCI. But there are really two types of biases here: nonresponse bias and measurement error bias[7]. Nonresponse bias is a mis-alignment of the establishment sample to the population caused by differential response rates. In many cases, nonresponse bias can be kept in check when a good nonresponse analysis is followed by effective nonresponse weighting adjustments. These adjustments can generally eliminate the largest portion of this bias successfully. We assume that both SOII and OSHA EDCI estimates will be put through this type of analysis and provided with appropriate nonresponse weighting adjustments. After this exercise is completed, it may be reasonable to assume nonresponse biases are not problematic (this does not mean that nonresponse adjustments are perfect, but that it will be acceptable to treat both sets of estimates as effectively unbiased).

The other type of bias that is harder to adjust for is measurement error bias. This is an error in the measurement provided to the government as a measure of the true degree of OIIs in the

[7] For example, errors in reporting industry and employment size classes would be measurement error.

establishment in the current time period. Suppose we write Y_{tika} as for example the true total number of DAFW or DJTR for state t, industry i, establishment size class k, establishment a for the survey year. Write $y_{tika}^{SOII\_SR}$ as the self-reported value of Y_{tika} from a SOII respondent (conditioning on their being sampled for SOII and responding to SOII). The measurement error in the self-report is $e_{tika}^{SOII\_SR} = y_{tika}^{SOII\_SR} - Y_{tika}$. The expectation is $E\left(e_{tika}^{SOII\_SR}\right)$. If there is a bias, it is likely to be a negative because it might be due to underreporting (not including cases on the OSHA forms that should be included), i.e., we might expect $E\left(e_{tika}^{SOII\_SR}\right) < 0$.

For OSHA EDCI measurement there is a corresponding $y_{tika}^{OSHA\_SR}$, with a corresponding $e_{tika}^{OSHA\_SR} = y_{tika}^{OSHA\_SR} - Y_{tika}$. The expectation is $E\left(e_{tika}^{OSHA\_SR}\right)$, and we might expect $E\left(e_{tika}^{OSHA\_SR}\right) < 0$ also.

Measuring the bias $E\left(e_{tika}^{SOII\_SR}\right)$ is very difficult without some type of enhanced followup. It may be that SOII estimates may have a systematic unmeasurable bias $E\left(e_{tika}^{SOII\_SR}\right) < 0$, that is a permanent part of the SOII estimation system. SOII in fact is estimating the OSHA log version of OII incidence, and not ultimately the true OII incidence. Thus, a reasonable goal is that the OSHA EDCI bias $E\left(e_{tika}^{OSHA\_SR}\right)$ is equal to $E\left(e_{tika}^{SOII\_SR}\right)$ and not instead less than $E\left(e_{tika}^{SOII\_SR}\right)$.

4.2 Comparison of Area-Level Estimates

We present two possible ways to check for measurement error bias. The first (described in this section) is to compare area-level estimates coming from OSHA EDCI and SOII for the same OII statistics and within the same domains. The second uses a direct linking of single establishments, and is described in the next section. The second way may be a much more efficient methodology, but it may not be possible especially once the sample design for SOII is altered not to overlap with OSHA EDCI data collection. Thus the area-level comparison may end up in some cases to be the only possibility. Area-level comparisons can be separated out into direct and indirect comparisons, as described in Sections 4.2.1 and 4.2.2 respectively.

4.2.1 Direct Comparison of Area-Level Estimates

The gold standard would be a direct comparison of SOII and OSHA EDCI estimates for a particular OII statistic and a particular domain. Both of these estimators should be weighted

estimators, fully adjusted for nonresponse, so that nonresponse bias is removed from the table as far as is feasible. Standard errors should be computed, and a standard error of the difference computed assuming independence between the estimates. In this case, the difference is an estimate of the bias, and a t-statistic can be computed for testing purposes. It may be difficult to interpret a large number of standardized differences due to the many estimates possible, but it may be possible to do a sign-test of the direction of the biases. If OSHA EDCI does in fact suffer from a relative measurement error bias, we expect it to be in one particular direction (specifically, a smaller estimate for OII prevalences than SOII). Thus, an unexpectedly large number of differences in one direction (standard difference significant or not) will be powerful evidence of bias.

Direct comparison will be possible in the early stages of this initiative when the SOII sample design has not been altered to avoid OSHA EDCI domains. Every OSHA EDCI Form 300A estimate will be matched by a SOII Form 300A estimate. A broad omnibus sign test can be done over all a large number of separate domains using a key prevalence characteristic such as total rate of DAFW. When the SOII sample design is altered to avoid overlap, the incidence of these direct comparisons may drop off considerably, and other methods may need to be utilized.

4.2.2 Indirect Comparison of Area-Level Estimates

One indirect comparison that may be common between OSHA EDCI and SOII is a comparison between OSHA EDCI Form 300A estimates and aggregations developed from SOII Form 300/301 estimates. The SOII Form 300/301 estimates can be aggregated up to match in principle some of the OSHA EDCI Form 300A estimates. This comparison may be fairly close to a direct comparison, and these types of comparisons will certainly be possible during the time period when OSHA EDCI is collected only for Form 300A, and SOII continues to be collected for a sample for Form 300/301.

It may be necessary to have even more indirect comparisons when sample design changes in the SOII result in no overlap with OSHA EDCI at all in particular domains. Another issue may be that SOII estimates may be affected in a hidden way due to the fact that establishments in particular domains are required to cooperate with OSHA EDCI. If they have underreported to OSHA because they realize their OSHA EDCI data is being published on a publicly transparent venue, then they may make their reports to SOII also consistent with the underreport sent to OSHA even though they know that SOII is confidential unlike OSHA EDCI. For example, they may base both reports

on the same logs. It is speculative at this point to say that this might take place, but it seems prudent to at least allow for the possibility. We know in general that differing degrees of transparency versus confidentiality can create a mode effect. If this mode effect has occurred, then this might bias the SOII estimate itself whereas it was not biased beforehand. It may be possible (though difficult) to pick this up by looking for breaks in the SOII time series that can be attributed to the onset of the OSHA EDCI.

Regression discontinuity methods is one methodology that searches for the potential mode effect breaks in the SOII time series. One of the original papers on regression discontinuity was Thistlewaite and Campbell (1960). A recent overview paper is Imbens (2016). Regression discontinuity works by working with a threshold for receiving a 'treatment' based on some kind of administrative procedure, regulation, etc. The original Thistlewaite and Campbell (1960) paper describes research on the effect on high-school students who receive a 'Certificate of Merit' from the National Merit Scholarship Program. This can affect their self-image and sense of future career possibilities as measured by later questionnaires. This Certificate of Merit to the student is awarded based on having an achievement score higher than a set threshold. Attitudes can be modeled as linear functions of achievement, and any discontinuity in the regression line is an indication of 'treatment effect': a psychological effect on the student of receiving this Certificate of Merit. It is important that the underlying regression function is actually linear (or at least locally linear), and there are no confounding factors. With these assumptions, the break in the regression line can successfully measure the treatment effect. The threshold due to its relative arbitrariness can function to separate into treatment and control like a classical randomized experiment (not perfectly, but reasonably well). These methods have been used widely in the past twenty years or so (see Imbens (2016)).

In this case, the regression discontinuity threshold might be the cutoff for establishment size for being required to participate in the OSHA EDCI initiative. These cutoffs are 20 employees for historically high-risk industries, and 250 employees for historically low-risk industries. An 'RD domain' can be defined around these cutoffs: say 15 to 50 employees for high-risk industries, and 150 to 500 employees for low-risk industries, straddling on either side the OSHA EDCI threshold (the cutoffs on either side should be set to provide relatively equal sample sizes below and above, and sufficient sample sizes for an analysis while remaining close to the threshold to justify local linearity). An analysis can be done using historical SOII data relating OII rates to stratum and establishment size to provide a baseline. Getting the right relationship to establishment size conditional on stratum is an important aspect (flat, linear, nonlinear). A further analysis can then be done on the same RD domain with the data under the aegis of the OSHA EDCI initiative. The

regression discontinuity analysis looks for breaks in the relationship in the joint SOII-OSHA EDCI estimates between establishments above and below the threshold (those who are part of the OSHA EDCI, and those who are not). Changes from the baseline could be from sources other than the EDCI, so this type of analysis can never be conclusive. Nevertheless, it certainly can provide a great deal of evidence where there is no longer a direct establishment overlap between SOII and OSHA EDCI. We recommend it should be done to confirm no mode shift.

4.3. Checking for Bias Through Unit-Level Checks

A more powerful way to check for measurement error bias is to check for differences in measurements between establishments reporting to OSHA EDCI and also reporting to SOII. A potential source of confounding is whether the fact that the establishment is participating in both (and is likely to know this) will result in a mode shift in the SOII result, as discussed above. This possibility has to be kept in mind.

For the early phase of the OSHA EDCI when SOII data collection is not yet altered, and assuming there is a set of reliably linked pairs of overlapping SOII and OSHA EDCI establishments, the bias analysis could compare the two sets of Form 300A types of counts. Ideally, the SOII and OSHA EDCI OII counts will completely agree, and this should be true for many if not most establishments. The incidents of nonagreement should be studied and understood. In particular, the major concern should be in systematic differences (positive or negative) between the SOII counts and the OSHA counts from the same establishment for the same year, and whether there are systematic differences by establishment characteristic (i.e., biases by establishment subgroups).

In a later phase of the OSHA EDCI when SOII data collection drops Form 300A data collection in domains covered by OSHA EDCI, the bias analysis could compare the two sets of Form 300A OSHA EDCI counts to the case level Form 300/301 data from the same establishment in SOII.

In order to carry out this micro-level analysis, linking will be necessary. Appendix B provides a discussion as to how best to do this linking.

Area-Level Models in the Presence of Nonnegligible OSHA EDCI Bias

5

In this section, we discuss a methodology for area-level composite estimation when the OSHA EDCI are found to have relative measurement error bias (or are at least suspected of possibly suffering from this bias). A simple composition or calibration procedure does not work in this case: an explicit strategy needs to be used to actually model the bias and adjust for it.

The recommended approach is based on those developed from Lohr and Brick (2012). Suppose domains $d = 1, \ldots, D$ correspond to the domains tik (see Section 5.1) for which there are both SOII and OSHA EDCI estimates. The idea is to develop a final set of estimated totals $\hat{Y}_{tik}$ (denoted for simplicity as $\hat{Y}_d$) across these domains $d = 1, \ldots, D$ with both types of estimates available. Suppose $\bar{y}_d$ is the weighted estimate over the sampled establishments for the establishment OII count (for a particular type of OII) divided by the total working hours represented for the establishments for domain d from SOII, and $\bar{x}_d$ is the corresponding weighted mean for domain d from OSHA. Then we assume as in Lohr and Brick (2012) the following model:

$$\begin{pmatrix} \bar{y}_d \\ \bar{x}_d \end{pmatrix} \sim N\left[\begin{pmatrix} \theta_d \\ \theta_d + \eta_d \end{pmatrix}, \sigma^2 \begin{pmatrix} n_{yd}^{-1} & 0 \\ 0 & n_{xd}^{-1} \end{pmatrix}\right]$$

The random effects parameter θ_d is assumed to be the actual population value. The random effects parameter η_d then measures the bias in the OSHA EDCI estimator in this domain. It should be noted that the bias in the OSHA EDCI estimator that is being modeled here is a measurement-error type bias and not a nonresponse-induced population distribution type of bias (as discussed in Section 4.1). This latter type of bias can be dealt with through nonresponse weighting adjustments, and these nonresponse-adjusted sampling weights should be included in the likelihood calculation (making the likelihood a 'pseudo-likelihood': see for example Pfeffermann (1993)). n_{yd} and n_{xd} are effective establishment sample sizes for domain d for SOII and OSHA EDCI respectively. The distribution of random effects is assumed to be:

$$\begin{pmatrix} \theta_d \\ \eta_d \end{pmatrix} \sim N\left[\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} a & b \\ b & c \end{pmatrix}^{-1}\right]$$

where μ_1 is the overall population mean, and μ_2 is the overall population bias (the relative bias of the OSHA EDCI estimates). The parameter a measures the degree of variability across domains in the population mean, and the parameter c measures the degree of variability across domains in the bias levels (small values of the parameters here correspond to large variability: they are precision parameters). The parameters μ_1, μ_2, a, b, c are estimated through maximum likelihood. σ^2 is treated as a fixed value and derived from design-based type calculations similar to Fay and Herriot (1979). From Lohr and Brick (2012), σ^2 is 'assumed known; in practice, a smoothed estimator from the design-based variances is used'. It should be noted that this maximum likelihood model is only fit over the domains that include OSHA EDCI estimators, a subset of the full universe of all SOII domains.

For this procedure to be practical in production, separate maximum likelihood estimation would not be done for every estimation item, every year. The model could be fit periodically, with the fitted models suggesting estimators that can be used in production. Lohr and Brick (2012) recommend this, suggesting (their Equation (12)):

$$\hat{\theta}_{d,1} = \hat{\lambda}_{d,1}\bar{y}_d + \left(1 - \hat{\lambda}_{d,1}\right)(\bar{x}_d - \hat{\mu}_2) \text{ with}$$

$$\hat{\lambda}_{d,1} = \frac{n_{xd}n_{yd} + \hat{c}\hat{\sigma}^2 n_{yd}}{n_{xd}n_{yd} + \hat{c}\hat{\sigma}^2\left(n_{xd} + n_{yd}\right)}$$

Note that $\hat{\sigma}^2$ is obtained based on the design variances (it is not a maximum likelihood estimator). $\hat{\mu}_2$ and $\hat{c}$ are maximum likelihood estimators based on the model, assuming implicitly that $a = 0$. Note that this assumption that $a = 0$ has the implication that the $\bar{y}_d$ estimates are used from SOII for each domain d without any shrinkage: this is done currently. Only the bias estimates are subject to shrinkage. Closed-form solutions can be obtained for $\hat{\mu}_2$ and $\hat{c}$ that can be used in a production setting. It is unlikely that this estimation procedure can be implemented simply by generating specialized weights for each establishment: the weights that would be applied to the non-domain establishments for the estimate for any particular domain would have to vary across domains. Each domain estimate for each estimation item would need to be generated using the composite estimator. But the same production algorithm should be usable for each domain and each estimation item (with different inputs).

A single set of establishment-level weights could be developed that would be sub-optimal composites for any given estimation item, but might have reasonable precision levels for many estimation items, and also be additive as domains are aggregated together and estimation items

combined. This single set of weights would be something akin to a 'minimax' set of weights: minimizing the maximum variance over a set of estimation items. Research would need to be done to see if this is feasible.

It should be noted also that the pseudo maximum likelihood estimator is finally producing an estimator of θ_d, which is a mean for the domain over establishments of OII incidence per working hour (for each establishment, the number of OII events of a particular type divided by the working hours). If one wants a domain estimate of the total number of OII incidents, then this θ_d estimate needs to be multiplied by an estimate of the mean over establishments of total working hours. The OSHA version of this latter estimate is less likely to suffer from bias, and can be composited by the simpler methods of Section 3.1.

It may also be possible to tweak the model and still maintain a fairly simple production-friendly estimator. For example, instead of a single 'grand' bias μ_2 one might have an ANOVA type model of separate stratum bias means $\mu_2(t), t = 1, \ldots, T$ within particular domains $t = 1, \ldots, T$.

The proposed way to determine an appropriate model is through either a macro- or a micro-analysis of the differences between OSHA-EDCI and SOII reports from the same establishments. The macro-analysis would be a direct analysis of domain estimates $\bar{y}_d$ and $\bar{x}_d$, finding a domain-specific structure for the bias (if one can be sufficiently discerned). The micro-analysis would follow a Section 4.3 type unit analysis of individual establishment bias differences. This analysis can only be done if there is a matching OSHA-EDCI and SOII report on the same establishment in the same time period (possibly OSHA-EDCI Form 300A vs. SOII Form 300A, or possibly OSHA-EDCI Form 300A vs. SOII Form 300/301 if SOII 300A data collection has been discontinued).

Variance estimation of the composite estimates can be accomplished by developing replicate weights for both the OSHA and the SOII estimates. Finite population corrections would need to be included explicitly. The OSHA replicates would be entirely based on nonresponse adjustments, as there is no sampling. OSHA and SOII would be assumed to be independent and the replicate weights should reflect this. With this replicate weights structure in place, the composition work can be fully replicated, and this should provide consistent estimators for the composite estimates.

6 Discussion

In the Literature Search Review and Methodology Report I for this research project we explored many advanced methodologies for dealing with relative measurement error bias in OSHA EDCI. These methods include area-level approaches beyond the one described in Section 5 above, and also include unit-level approaches that put together micro-data from SOII and OSHA EDCI and combine them directly. The complexity of these unit-level models makes it unlikely that these procedures can realistically be part of a permanent SOII production cycle.

We were unable to find documented evidence that any of the more advanced methods for combining surveys (Sections 5.2 and 5.3, Sections 6.2 and 6.3 of Methodology Report I; most of the more advanced approaches in the literature search report) are actually being used in a regular production mode in US Federal Statistical agencies (or anywhere else). These methods seem to be still in a research mode rather than a production mode. One exception is the NHIS-BRFSS cancer risk factor and treatment prevalences as developed in Raghunathan et al. (2007). County-level prevalence estimates based on combining NHIS and BRFSS using the advanced models in Raghunathan et al. (2007) can be downloaded from a **https://sae.cancer.gov/nhis-brfss** website as described in Appendix C. Even in this case, the most recent years available are 2008-2010, so it appears there is a considerable lag in preparing these estimates (if they are still being done). There is another longstanding (over several decades now) small-area estimation program using advanced models in a regular production mode: SAIPE (Small Area Income and Poverty Estimates) carried out by the US Bureau of the Census, and these estimates are available in a regular production mode (one can link with **https://www.census.gov/programs-surveys/saipe**). Estimates are available for states, counties, and geographic school districts. These are small-area models only: they do not combine multiple surveys but rely on the American Community Survey estimates exclusively. Nevertheless, it is an example of the use of advanced empirical Bayes type models used in a regular production setting.

References

Cochran, W. G. (1977). *Sampling Techniques, 3rd ed.* New York: John Wiley & Sons.

Christen, P. (2012). *Data Matching*. New York: Springer.

Dever, J. A., and Valliant, R. (2010). "A comparison of variance estimators for poststratification to estimated control totals". *Survey Methodology 36, 1*, 45-56.

Fellegi, I. P., Sunter, A. B. (1969). "A theory for record linkage". *Journal of the American Statistical Association 64*, 1183-1210.

Francisco, C. A., and Fuller, W. A. (1991). "Quantile estimation with a complex survey design". *The Annals of Statistics 19, 1*, 454-469.

Imbens, G. (2016). "Regression discontinuity designs in the econometric literature". *Observational Studies 2*, 147-155.

Kish, L. (1992). "Weighting for unequal pi". *Journal of Official Statistics 8 (2)*, 183-200.

Lohr, S. L. (2011). "Alternative survey sample designs: Sampling with multiple overlapping frames." *Survey Methodology 37, 2,* 197-213.

Lohr, S. L., and J. M. Brick (2012). "Blending domain estimates from two victimization surveys with possible bias." *The Canadian Journal of Statistics 40, 4,* 679-696.

Oh, H. L., and Scheuren, F. J. (1983). "Weighting adjustment for unit nonresponse", in W. G. Madow, I. Olkin, D. B. Rubin (eds.), *Incomplete Data in Sample Surveys, Vol. 2,* New York: Academic Press, pp. 143-184.

Pfeffermann, D. (1993). "The role of sampling weights when modeling survey data." *International Statistical Review*: 317-337.

Pierce, B. (2017). "Prospects for combining survey and non-survey data sources to improve estimated counts of certain work-related injuries". Private communication: BLS technical report.

Raghunathan, T. E., Xie, D., Schenker, N., Parsons, V. L., David, W. W., Dodd, K. W., and E. J. Feuer (2007). "Combining information from two surveys to estimate county-level prevalence rates of cancer risk factors and screening". *Journal of the American Statistical Association 102,* 474-486.

Selby, P. N., Burdette, T. M., and E. M. Huband (2008). Overview of the Survey of Occupational Injuries and Illnesses sample design and estimation methodology. Available at https://www.bls.gov/osmr/pdf/st080120.pdf.

Thistlewaite, D. L., and Campbell, D. T. (1960). "Regression-discontinuity analysis: An alternative to the ex-post facto experiment." *Journal of Educational Psychology 51 (6)*, 309-317.

Valliant, R., Dever, J. A., and Kreuter, F. (2013). *Practical Tools for Designing and Weighting Survey Samples*. New York: Springer.

Appendix A
Compositing Population Quantiles

Francisco and Fuller (1991) provides a widely cited paper regarding the estimation of population quantiles in the context of stratified cluster sampling. Their approach is to define the population distribution function $F(x)$ at each point x in the population support set, estimate the population distribution function using the general stratified cluster design unbiased estimator, and then to invert it for each desired quantile q, finding the value x such that $F(x) = q$).

The approach below is designed for the quartile estimates (1st quartile $q = 0.25$; median $q = 0.5$; 3rd quartile $q = 0.75$) of different case types (all recordable cases, DAFW cases, DJTR cases, etc.). The domains d are for example NAICS industry classes crossed with various establishment size categories appropriate for a particular industry. We will assume for the sake of simplicity in this section that the domain d nests fully within one establishment class stratum $k = 1,..,4$, and fully within one industry $i = 1, \dots, I$ (either within the low-risk industry stratum $I(L)$ or the high-risk industry stratum $I(H)$). With this simplifying assumption the domain d is either entirely covered by SOII only (in which case there is no need for theoretical development), or is entirely covered by both SOII and OSHA. The more complicated case of non-nesting simply involves breaking the domain into two subdomains: one of which is SOII only, and the other is completely SOII and OSHA. Write $j = 1, \dots, J_d$ as the set of establishments within the domain. Following the Francisco and Fuller approach, write the population distribution function for domain d at the point x as

$$F_d(x) = \frac{1}{N_d} \sum_{j=1}^{J_d} I\left[Y_{dj} \leq x\right]$$

where N_d is the total number of establishments in domain d, $I[\dots]$ is an indicator function which is 1 if the argument is true, and is 0 otherwise, with $I\left[Y_{dj} \leq x\right]$ equal to 1 if $Y_{dj} \leq x$ and is 0 otherwise (thus equalling 1 if the case count for the establishment is less than x, and equal to 0 otherwise). Write the quantile $Q_d(q)$ for domain d, quantile q as

$$Q_d(q) = \min_x \{F_d(x) \geq q\}$$

The recommended approach to estimation based on Francisco and Fuller is to generate composite estimators of the distribution function for values of x, and then invert this composite distribution

function to generate the quantile. For each domain d, case type y and x value we have the following. We are assuming that the domain d is a domain for which there are SOII and OSHA estimates, and that the OSHA estimate covers the entire domain (rather than only a subdomain). Let w_d^{SOII} be the composite weight factor for domain d allocated to SOII. Then $1 - w_d^{SOII}$ is the composite weight factor allocated to OSHA. Write ew_{dj}^{SOII} as the establishment's weight in the SOII sample and write ew_{dj}^{OSHA} as the establishment's weight in the OSHA sample. One possible composite estimator (the 'separate composite first, then invert') for the distribution function evaluated at x is:

$$\hat{F}_d^{(1)}(x) = w_d^{SOII} \frac{\sum_{j=1}^{J_d} ew_{dj}^{SOII} I[Y_{dj} \le x]}{\sum_{j=1}^{J_d} ew_{dj}^{SOII}} + (1 - w_d^{SOII}) \frac{\sum_{j=1}^{J_d} ew_{dj}^{OSHA} I[Y_{dj} \le x]}{\sum_{j=1}^{J_d} ew_{dj}^{OSHA}}$$

The corresponding quantile estimator is

$$\hat{Q}_d^{(1)}(q) = \min_x \left\{ \hat{F}_d^{(1)}(x) \ge q \right\}$$

A second possible composite estimator (the 'combined composite first, then invert') for the distribution function evaluated at x is:

$$\hat{F}_d^{(2)}(x) = \frac{w_d^{SOII} \sum_{j=1}^{J_d} ew_{dj}^{SOII} I[Y_{dj} \le x] + (1 - w_d^{SOII}) \sum_{j=1}^{J_d} ew_{dj}^{OSHA} I[Y_{dj} \le x]}{w_d^{SOII} \sum_{j=1}^{J_d} ew_{dj}^{SOII} + (1 - w_d^{SOII}) \sum_{j=1}^{J_d} ew_{dj}^{OSHA}}$$

with corresponding quantile estimator

$$\hat{Q}_d^{(2)}(q) = \min_x \left\{ \hat{F}_d^{(2)}(x) \ge q \right\}$$

A third alternative is to 'invert before compositing'. This is outside the Francisco-Fuller framework. Generate two separate quantile estimators based only on SOII and OSHA respectively, and then composite the quantile estimators:

$$\hat{F}_d^{SOII}(x) = \frac{\sum_{j=1}^{J_d} ew_{dj}^{SOII} I[Y_{dj} \le x]}{\sum_{j=1}^{J_d} ew_{dj}^{SOII}} \qquad \hat{F}_d^{OSHA}(x) = \frac{\sum_{j=1}^{J_d} ew_{dj}^{OSHA} I[Y_{dj} \le x]}{\sum_{j=1}^{J_d} ew_{dj}^{OSHA}}$$

$$\hat{Q}_d^{SOII}(q) = \min_x \left\{ \hat{F}_d^{SOII}(x) \ge q \right\} \qquad \hat{Q}_d^{OSHA}(q) = \min_x \left\{ \hat{F}_d^{OSHA}(x) \ge q \right\}$$

$$\hat{Q}_d^{(3)}(q) = w_d^{SOII} \hat{Q}_d^{SOII}(q) + (1 - w_d^{SOII}) \hat{Q}_d^{OSHA}(q)$$

Which of the three estimators will be best? The differences between 'separate composite, then invert' and 'combined composite, then invert' are likely to be small. Cochran (1977), Section 6.12 analyzes the differences between the combined ratio estimator and the separate ratio estimator in the context of stratified random sampling, and finds limited differences. Generally the separate ratio estimator is slightly better unless the denominator has a high CV. This would favor slightly the 'separate composite, then invert' version. What of the 'invert before compositing' version? As long as the two quantiles $\hat{Q}_d^{SOII}(q)$ and $\hat{Q}_d^{OSHA}(q)$ are both sufficiently stable, compositing them should lead to a good result. A further comparison would likely require empirical comparisons and possibly simulation studies.

Appendix B
Probabilistic Linkage

Probabilistic linkage between SOII and OSHA EDCI records will likely be an important part of the methodology for checking whether or not there is relative measurement error bias in the OHSA EDCI estimates.

If the OSHA EDCI establishment electronic data collection includes the Employment Identification Number (EIN), then linkage between OSHA EDCI and SOII records may become deterministic, or near-deterministic. A false linkage can only occur if there is an error in the transcription of the EIN in either OSHA EDCI or SOII. Even assuming linkage though by EIN, it may be useful to have the full machinery of linkage modeling to allow for the possibility of misspecified EIN. The mismatching on other fields that would be apparent if the EIN was misspecified and the link was actually false would be very clear, and the probability of a correct link (temporarily setting aside the EIN) would register as extremely low. These cases could be forwarded then for manual review, and the EIN could be duly corrected. This procedure is likely to be valuable under the scenario that EIN is available for deterministic linking.

The variables that should be compared between SOII and OSHA EDCI for linking include the following:

- Company name;
- Company street address;
- Company city;
- Company state;
- Company zip code;
- SIC or NAICS industry code;
- Annual average number of employees;
- Total hours worked by all employees;
- Total number of deaths from OII;

- Total number of cases with days away from work;
- Total number of cases with job transfer or restriction; and
- Total number of injuries by OII type (e.g., poisonings).

The company name and address should be standardized first using standardization software for preparing names and addresses for linking. There are a variety of such standardization software packages available.

Suppose we write the twelve characteristics above coming from the two reports $X_{ika}(SOII)$ and $X_{ikb}(SOII)$, where $i = 1, \ldots, 12$ represents the characteristic, $k = 1, \ldots, K$ represents K 'blocks', a represents establishment a in the SOII sample in block k and b represents establishment b in the OSHA EDCI data collection in block k (establishment ka and kb may or may not be pairs). The blocks are matching sets of establishments in each sample which are assumed to contain all true pairs (i.e., we assume there are no pairs going across blocks). Blocking is necessary as otherwise the number of potential pairs will become far too large. It is important to select the blocks carefully: that is an important analysis topic. Any true pairs (a, b) that are in different specified blocks k_1 and k_2 will end up being lost. See for example Christen (2012), Section 4.2.

Define distance functions $\gamma_{ikab}\left(X_{ika}(SOII), X_{ikb}(OSHA)\right)$ which are zero if the two X values match completely, are positive if there is a discrepancy, and are larger if the discrepancy is larger. For the company name, city, and street address, the distance function would be some type of count of discrepancies between the text fields, ranging from zero for the standardized text fields being identical, to a large value for text fields that are entirely different. Section 5 in Christen (2012) provides an extensive listing of current approaches for evaluating text string accuracy. State would be a simpler distance: either the state abbreviations are the same or they are different. Zip code might be based on a hierarchy: differences between the last four characters or not of a nine-character zip, differences between the fifth characters in a five-character zip, etc. These methods or others would need to be tested in this context (based on training sets and manual quality checks) to develop the most accurate methodology.

A distance function for SIC or NAICS code would be based as zip code on a hierarchy of left-to-right characters (first character to the left is at the highest level of the hierarchy: last character to the right is the lowest level). If one report has SIC code and the other NAICS, then a linking table between SIC and NAICS would need to be utilized. Any distance code would be specially developed

for these industry codes (this is beyond the scope of this methodology report: requires economists' input).

For the six numerical totals, the distance function can be an absolute number (the absolute value of the difference), or the absolute value of the log of the ratio between the two totals (a ratio equaling 1 corresponding to complete concordance). There are other similar distances (see for example Section 5.12 of Christen (2012)).

We standardize all of these twelve distances to fall within the unit interval [0,1], with 0 corresponds to perfect match, and 1 complete nonmatch.

Modifying somewhat the Fellegi-Sunter notation (Fellegi and Sunter 1969), we define $M(k)$ as all pairs (ka, kb) in block k where there is a match between the records (they are the same establishment). We define $U(k)$ as all pairs (ka, kb) in block k where there is not a match between the records (they are different establishments). Define γ_{kab} as a 12-vector of distance functions for the pair (ka, kb):

$$\gamma_{kab} = \{\gamma_{ikab}\}_{i=1,\ldots,12}$$

A perfect match corresponds to a γ_{kab} vector equal to a vector of all zeroes. A complete nonmatch corresponds to a γ_{kab} vector with sizeable positive values (in the extreme case, all 1's). A match that has some errors might have many zero components and a few positive components not much larger than 0.

In many applications of this methodology, the assumption is made that the components of the gamma vector are independent: i.e.,

$$Pr\left\{\gamma_{kab} \in \prod_{i=1}^{12} \Gamma_i\right\} = \prod_{i=1}^{12} Pr\{\gamma_{ikab} \in \Gamma_i\}$$

Under this independence assumption, the probability of 12-dimensionsal cube of intervals is the product of the probability of the individual intervals. But in this circumstance, this assumption is probably too strong. We expect correlations in the matching of address, city, and state, and between the various totals.

The best way to proceed might be to prepare a 'training data set' consisting of pairs (ka, kb) from a number of blocks which have their pairings assigned manually by trained personnel, so that the right answer can be discerned. One also then has a large set of non-matches naturally as well (all of the other pairings between records on the two data sets). Thus one knows $M(k)$ and $U(k)$ completely for the training data set blocks.

With this information, one can find clusters Γ_c which correspond to areas of homogeneous propensity to be in $M(k)$ or $U(k)$. There will be clusters Γ_c with very high propensity $Pr\{kab \in M(k) | \gamma_{kab} \in \Gamma_c\}$ close to 1. These are clusters that include the zero vector (all distances zero). There will be clusters Γ_c with very low propensity $Pr\{kab \in M(k) | \gamma_{kab} \in \Gamma_c\}$ close to 0. These are clusters which have mostly positive distance values. There may or may not be clusters with some components close to zero distance and others with positive distance, with middling propensities (not close to either 0 or 1). This clustering task can be done using cluster-generating or tree-generating software. There are a variety of software possibilities. This will generate in principle a mutually exclusive and exhaustive set of cells Γ_c, $c = 1, \ldots, C$ that partition the full universe (the twelve-dimensional cube $[0,1]^{12}$) and for which we can assume $Pr\{kab \in M(k)\}$ is constant within each cell Γ_c.

With these developed cells Γ_c, the production data set can be processed. For each pair (ka, kb) in the production data set (within block k) we can assign standardized distances $\hat{\gamma}_{kab}$, and assign an estimated probability $\hat{p}_{kab}(M(k))$ based on the cell $\hat{\Gamma}_c$ that contains $\hat{\gamma}_{kab}$, with $\hat{p}_{kab}\big(U(k)\big) = 1 - \hat{p}_{kab}\big(M(k)\big)$.

A decision rule can be developed for assigning (ka, kb) as a matched pair or not:

$$\begin{array}{cc} \hat{p}_{kab}\big(M(k)\big) > p_u & \text{matched pair} \\ p_l < \hat{p}_{kab}\big(M(k)\big) < p_u & \text{subject to manual review} \\ \hat{p}_{kab}\big(M(k)\big) \leq p_l & \text{no} - \text{match pair} \end{array}$$

The 'matched pairs' are assumed then to be matches and are treated as such in further analyses. The 'no-match pairs' are assumed to not be matches. The 'subject to manual review' group has an uncertain assignment and should be checked manually. The cutoffs for the intervals should be set based on an empirical review of numbers of false matches and false non-matches (following Fellegi and Sunter (1969)). One wants to minimize the prevalence of the numbers of false matches and false non-matches, and also minimize the percentage of cases sent to manual review. The decision rule can be tweaked based on manual quality checks of pairs assigned to both the matched pair and no-match pair.