Statistical Disclosure Limitation and Total Survey Error (Proposal)

Evan Totty *U.S. Census Bureau*

Thor Watson *U.S. Census Bureau*

September 12, 2022

1 Introduction

The purpose of this proposed project is to expand the total survey error framework to include error from statistical disclosure limitation (SDL) and quantify this error relative to other types of survey error. In Section 2 of this document, we present the extension of the total survey error framework to include error from SDL. In Section 3, we discuss our plans to use this framework to quantify SDL error relative to other types of survey error found in U.S. Census Bureau survey data.

The total survey error framework quantifies traditional sources of survey error such as coverage error, measurement error, and non-response error (Biemer, 2010; Groves and Lyberg, 2010; Meyer and Mittag, 2021*b*). Assessing the relative size and impact of different types of error is useful for statistical agencies when determining where to allocate scarce resources available for improving survey quality. It is also useful for researchers when determining whether certain data are fit for use for a given purpose. Quantifying each type of survey error can be challenging given data requirements, but linked survey and administrative data make this possible (Meyer and Mittag, 2021*b*). We extend both the framework and the empirical analysis in recent work by Meyer and Mittag (2021*b*) to account for error that arises due to the application of SDL to survey data.

Statistical agencies such as the Census Bureau use SDL mechanisms to reduce the risk of privacy loss for survey and census respondents. SDL is a post-processing step that occurs between when the data is collected and when it is released to the pubic. During this step, any number of SDL techniques may be applied, including geographical and categorical suppression, rounding, top-coding, data swapping, and noise infusion. This step is applied to both microdata and tabulated statistics. There is an inherent trade off between data privacy and data accuracy involved in the application of SDL: while these techniques reduce privacy loss risk, they also reduce the utility of the data. Suppression reduces utility by removing information altogether, while rounding, top-coding,

swapping, and noise infusion reduce utility by introducing uncertainty into the information. Abowd and Schmutte (2015) provide more discussion on the impact of SDL techniques on economic analysis.

Interest in the application and effect of SDL has grown in recent years, due in part to the Census Bureau's disclosure avoidance modernization efforts and use of a new disclosure avoidance system for the 2020 Decennial Census that featured differential privacy (Garfinkel, Abowd and Powazek, 2018; Abowd and Schmutte, 2019; Abowd et al., 2019; Jarmin, 2019; Ruggles et al., 2019; Cohen et al., 2021; Kenny et al., 2021; Hotz et al., 2022; Santos-Lozada, Howard and Verdery, 2022; Winkler et al., 2022). The use of differential privacy for the 2020 Decennial Census is part of a larger disclosure avoidance modernization effort underway at the Census Bureau (Abowd et al., 2020). Advances in computer science and computing resources combined with the growth in access to individual-level data from other sources in the age of "big data" has increased the risk of reconstruction and re-identification attacks on Census Bureau data (Abowd, 2016). These rising risks were the impetus for the disclosure avoidance modernization efforts implemented by the Census Bureau, and these risks motivate interest in further expanding the use of differential privacy and synthetic data at the Census Bureau (Abowd et al., 2020).

Importantly, modernization in how SDL is applied to statistical data products to reduce privacy loss risk must be combined with modernization in how we assess the impact of SDL on data accuracy. Trading off data accuracy and data privacy requires a diverse set of statistics for both accuracy and privacy. Furthermore, rising survey non-response rates and measurement error have already led researchers to conclude that household surveys are in a time of "crisis" (Meyer, Mok and Sullivan, 2015). This crisis partly motivated recent work quantifying several components of total survey error (Meyer and Mittag, 2019, 2021*a*,*b*; Meyer, Mittag and George, 2022). The authors show that measurement error is larger than both coverage error and non-response error in the datasets and variables they study. But their results do not quantify error that arises due to SDL error or post-processing error more generally.[1]

The key intuition for our analysis is that SDL techniques can be thought of as intentionally introducing error into the microdata or tabulated statistics. Furthermore, error due to SDL are only one of several types of errors that can exist in survey data. In order to fully understand the impact of SDL on data privacy and accuracy, we must calculate the amount and effect of SDL error compared to other types of survey error and conditional on other types of survey error. SDL error and other types of survey error may interact in ways that need to be understood. SDL error could cancel out other types of survey error in some cases or exacerbate it in others. We begin to fill in this gap by performing two specific analyses: (1) we expand the total survey error framework to include SDL error; (2) we quantify SDL error relative to coverage, measurement, and non-response error for a particular set of datasets and variables.

[1] Meyer and Mittag (2021*b*) mention the idea of incorporating post-processing error on page 300.

Our work contributes to a growing literature on the economic analysis of SDL. Abowd and Schmutte (2015) present econometric theory for the impact of SDL on various statistical techniques commonly employed in economic analysis. Abowd and Schmutte (2019) present the decision of how much SDL error to introduce as a resource allocation problem that weighs the demand for privacy against the demand for accurate statistics. Neither of the preceding analyses directly quantify the impact of error due to SDL relative to, or conditional on, other types of survey error. We argue that this contribution is an important one. Statistical agencies must make decisions about how to weigh and trade off accuracy versus privacy on a regular basis. The theory and practice of how to trade off data accuracy and data privacy is underdeveloped. But technical staff at statistical agencies have experience quantifying, improving, and trading off other types of survey error. Similarly, while many data users are not accustomed to accounting for SDL error in their analyses, they are used to accounting for traditional types of survey error. Quantifying SDL error relative to other types of survey error will provide statistical agencies and data users with important context for assessing the impact of SDL error.

The remainder of the paper is organized as follows. Section 2 presents the total survey error framework, including the extension to error from SDL. Section 3 discusses our plans for applying the results in Section 2 to data available at the Census Bureau.

2 Total Survey Error Framework

We begin with the framework from Meyer and Mittag (2021*b*) that defines total survey error based on combined survey and administrative data. Our contribution to the framework is that we add the concept of survey data before and after post-processing for disclosure avoidance, which in turn allows us to add SDL error to the framework. Otherwise, the framework, derivations, and assumptions all follow directly from their work. We summarize the framework below, including how to decompose it into coverage, measurement, non-response, and SDL error. Additional details, such as further decomposing coverage, measurement, and non-response into additional sub-components, can be found in Meyer and Mittag (2021*b*).

2.1 Total Survey Error Using Combined Data

The survey data is a sample of the population, while the administrative data contains information for the entire population. Let P denote the administrative population that the survey is intended to represent and S the population that the survey actually represents. The survey contains person final weights, w_i^f, that sum to the population size assumed by the survey, P^S. Let L denote the sub-set of survey observations that can be linked to the administrative data and assume that L can be re-weighted using $\hat{w}_i$ to account for missing linkages so that it is representative of S. Finally, let $r_i \in \{0, 1\}$ indicate whether individual

i responded to the survey.

We are interested in estimating a parameter μ of a vector of random variables X, where x_i is the value of X for individual i. To fix ideas, consider an application to wage and salary data, such that x_i will be an individual's wage and salary income and μ the average wage and salary income for the population. One could also analyze many other topics and statistics. For example, μ could be a measure of variance. One could also analyze sub-population sizes, in which case x_i would indicate whether individual i belongs to a particular group and μ would measure the group size.

There are up to three measures of x_i for each individual: x_i^A, which comes from the administrative records; x_i^S, which comes from the original survey data; and $\tilde{x}_i^S$, which comes from the post-processed survey data with SDL that is released to the public.[2] Only individuals who can be linked, $i \in L$, have a x_i^A measure. Furthermore, let $\hat{\mu}^A$ be the population average for the administrative records and $\tilde{\mu}^S$ be the weighted survey estimate on the post-processed data. We assume that the administrative records are accurate, meaning that there is no error in administrative record amounts for continuous variables and that individuals who exist in S but not A must be a false positive error in the survey data.[3]

For a given target parameter μ, the total survey error is the difference between the true population parameter and the survey estimate. We can estimate total survey error as the difference between the administrative data population average and the survey weighted estimate:

$$\hat{\varepsilon}_{TSE} = \frac{1}{\sum \hat{w}_i w_i^f} \left[\sum \hat{w}_i w_i^f \tilde{x}_i^S \right] - \hat{\mu}^A. \tag{1}$$

The first part of the expression is the (post-processed) survey weighted estimate of the target parameter. This is estimated from the linked sample, L, using both the final survey weights and the weight adjustments to account for the missing linkages. The second part of the expression is the population average for the entire administrative data. Note that some records may need to be excluded from the administrative data in order for it to represent the survey target. For example, if the survey excludes group quarters, then individuals

[2]We assume there is a person-level link between the original and post-processed data. That is, each record in the post-processed data corresponds to a particular record in the original survey data. This does not have to be the case. For example, when data are synthesized for privacy, they can be modeled in such a way that a synthetic record does not correspond directly to any underlying confidential record. However, many disclosure avoidance techniques do have a one-to-one link between pre- and post-processed records. This includes legacy methods such as swapping, top-coding, and rounding. Newer methods such as data synthesis may also include a one-to-one mapping, in particular those that are partially synthetic, which has been a common use of synthetic data to date (Hawala, 2008; Benedetto, Stanley and Totty, 2018).

[3]We acknowledge that assuming administrative records do not contain errors is incorrect, as it is well documented that they can also contain their own errors [see, e.g., Abowd and Stinson (2013)]. Our results should therefore be viewed with the caveat that we may falsely assign some administrative records error to survey error. However, we are following Meyer and Mittag (2021*b*) in making this assumption. We leave work that accounts for administrative records error when evaluating survey data error for future work.

living in group quarters should be removed from the calculation on the administrative data. Furthermore, administrative data that are not linkable should be removed because the survey data cannot cover them.

2.2 Decomposing Total Survey Error

We now decompose total survey error from equation (1) into coverage error, measurement error, non-response error, and SDL error. To derive the decomposition, replace $\tilde{x}_i^S$ in equation (1) with the equivalent expression $x_i^A + ((1 - r_i) + r_i)(x_i^S - x_i^A) + (\tilde{x}_i^S - x_i^S)$. This yields the following equation:

$$\begin{aligned}\hat{\varepsilon}_{TSE} =& \frac{1}{\sum \hat{w}_i w_i^f} \left[\sum \hat{w}_i w_i^f x_i^A\right] - \hat{\mu}^A \\ &+ \frac{1}{\sum \hat{w}_i w_i^f} \left[\sum \hat{w}_i w_i^f (1 - r_i)(x_i^S - x_i^A)\right] \\ &+ \frac{1}{\sum \hat{w}_i w_i^f} \left[\sum \hat{w}_i w_i^f (r_i)(x_i^S - x_i^A)\right] \\ &+ \frac{1}{\sum \hat{w}_i w_i^f} \left[\sum \hat{w}_i w_i^f (\tilde{x}_i^S - x_i^S)\right] \\ &= \hat{\varepsilon}_{GCE} + \hat{\varepsilon}_{INRE} + \hat{\varepsilon}_{ME} + \hat{\varepsilon}_{SDLE}\end{aligned} \tag{2}$$

The first term is generalized coverage error ($\hat{\varepsilon}_{GCE}$), which is the difference between the population average from the administrative data and the weighted survey total when the survey responses are replaced with administrative data values for the linked records. The second term is item non-response error ($\hat{\varepsilon}_{INRE}$), which is the weighted difference between the imputed survey response for individuals who did not respond and their administrative data value. The third term is measurement error ($\hat{\varepsilon}_{ME}$), which is the weighted difference between the survey response for individuals who did respond and their administrative data value. Finally, the fourth term is SDL error ($\hat{\varepsilon}_{SDLE}$), which is the weighted difference between the individual's response on the post-processed and original survey data.

Generalized coverage error, item non-response error, and measurement error are all a function of the difference between the administrative records and the original survey data, rather than the post-processed survey data, because these error components are all part of the survey design and implementation that occur before post-processing. SDL error is a function of the difference between the post-processed survey data and the original survey data, rather than the administrative records, because SDL is applied to the original survey data with the intent of minimizing the impact on statistics derived from the survey subject to some disclosure avoidance constraint.

Meyer and Mittag (2021*b*) further decompose $\hat{\varepsilon}_{GCE}$, $\hat{\varepsilon}_{INRE}$, and $\hat{\varepsilon}_{ME}$ into misclassification errors (false positives and false negatives) and errors in amounts. We can perform the same decomposition for $\hat{\varepsilon}_{SDLE}$. Let a false positive be an individual who reports a non-zero amount for wage and salary income in the post-

processed survey but not in the original survey, $FP = \{i \in L \& x_i^S = 0 \& \tilde{x}_i^S \neq 0\}$; a false negative be an individual who reports a non-zero amount in the original survey but not in the post-processed survey, $FN = \{i \in L \,\&\, x_i^S \neq 0 \,\&\, \tilde{x}_i^S = 0\}$; and errors in amounts be differences in reported amounts for those are who correctly classified as having non-zero amounts in both datasets, $CC = \{i \in L \,\&\, x_i^S \neq 0 \,\&\, \tilde{x}_i^S \neq 0\}$. Now we can further decompose SDL error into error from each of these components:

$$\hat{\varepsilon}_{SDLE} = \hat{\varepsilon}_{SDLE}^{FP} + \hat{\varepsilon}_{SDLE}^{FN} + \hat{\varepsilon}_{SDLE}^{Amount}, \tag{3}$$

where

$$\hat{\varepsilon}_{SDLE}^{FP} = \frac{1}{\sum \hat{w}_i w_i^f} \left[\sum_{i \in FP} \hat{w}_i w_i^f \tilde{x}_i^S \right] \tag{4a}$$

$$\hat{\varepsilon}_{SDLE}^{FN} = -\frac{1}{\sum \hat{w}_i w_i^f} \left[\sum_{i \in FN} \hat{w}_i w_i^f x_i^S \right] \tag{4b}$$

$$\hat{\varepsilon}_{SDLE}^{Amount} = \frac{1}{\sum \hat{w}_i w_i^f} \left[\sum_{i \in CC} \hat{w}_i w_i^f (\tilde{x}_i^S - x_i^S) \right]. \tag{4c}$$

Because all of the above calculations are done on the linked sample, a key assumption of this approach is that the linked sample is representative of its target after inverse probability re-weighting.[4] That is, $\mathbb{E}(x_i^S | i \in L, Z_i) = \mathbb{E}(x_i^S | Z_i)$, $\mathbb{E}(x_i^A | i \in L, Z_i) = \mathbb{E}(x_i^A | Z_i)$, and $\mathbb{E}(\tilde{x}_i^S | i \in L, Z_i) = \mathbb{E}(\tilde{x}_i^S | Z_i)$. Each of these assumptions is testable. Alternatively, SDL error and its components could be estimated based on the full survey sample rather than the linked sample. This would remove the threat that linkage status may be correlated with the variable of interest even after adjusting the weights, i.e., $\mathbb{E}(x_i^S | i \in L, Z_i) \neq \mathbb{E}(x_i^S | Z_i)$. The advantage of using the linked sample is that it holds the sample constant across all the different types of survey error so that the relative error amounts are comparable. As a robustness check we can also compute the SDL error components based on the full survey sample.

3 Data

In this section we discuss our plans for applying the total survey error framework described above to data from the Census Bureau. The Census Bureau has numerous high-profile surveys, such as the American Community Survey (ACS), Current Population Survey (CPS), and Survey of Income and Program Participation (SIPP). The Census Bureau also has many administrative records datasets via inter-agency agreements that can be used to assess and improve the

[4] We adjust the weights using inverse probability weighting. We specify a list of covariates, Z_i, that we think may be correlated with whether an individual is in the survey and then regress linkage status on these covariates for all individuals. The predicted linkage probability from this model is our weight adjustment, $\hat{w}_i$.

quality of its surveys and operations. Among them are the universe of 1040, W-2, and 1099 records from the Internal Revenue Service (IRS) as well as program participation information for linked respondents from the Social Security Administration (SSA). The Census Bureau also has population-level demographic information from the Decennial Census and a combination of administrative records sources (Ennis et al., 2018).

We intend to use the IRS universe records to study the total survey error components in survey-reported wage and salary income, retirement income, and self-employment. We intend to use the SSA data to study error components in social security disability, retirement, and aged spouse benefits. We intend to use the population demographic information from decennial censuses and administrative records to study error components in population sizes by race, Hispanic status, sex, and age.

We have the ability to study the impact of both legacy SDL methods, such as swapping, and modern approaches such as synthesis. The Census Bureau maintains versions of survey data before and after post-processing for disclosure avoidance that we can use for analyzing the impact of legacy approaches that have been used in the past. The Census Bureau has also developed a data synthesizer that we can use to produce fully synthetic or partially synthetic versions of surveys in order to study the impact of data synthesis.

4 Summary

Rising demand for data coupled with rising reconstruction and re-identification risk presents a challenge for statistical agencies such as the Census Bureau. Agencies have long used SDL to protect respondent information, but legacy methods are now seen as insufficient given the increasing sophistication of privacy attacks. Regardless of the SDL method, careful attention must be paid to the impact on usability of the data. This requires a holistic approach to survey error that moves beyond simply comparing post-processed data with SDL protection to pre-processed data without SDL protection. Rather, we must understand the impact of SDL protection relative to, and conditional on, other types of survey data. Our goal for this project is to provide an important step in that direction by quantifying and comparing error from SDL to error from other sources using the total survey error framework.

References

Abowd, John M. 2016. "How Will Statistical Agencies Operate When All Data Are Private?" *Journal of Privacy and Confidentiality*, 7(3).

Abowd, John M., and Ian M. Schmutte. 2015. "Economic Analysis and Statistical Disclosure Limitation." Brookings Papers on Economic Activity.

Abowd, John M., and Ian M. Schmutte. 2019. "An Economic Analysis of Privacy Protection and Statistical Accuracy as Social Choices." *American Economic Review*, 109(1): 171–202.

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

Abowd, John M., Gary L. Benedetto, Simson L. Garfinkel, Scot A. Dahl, Aref N. Dajani, Matthew Graham, Michael B. Hawes, Vishesh Karwa, Daniel Kifer, Hang Kim, Philip Leclerc, Ashwin Machanavajjhala, Jerome P. Reiter, Rolando Rodriguez, Ian M. Schmutte, William N. Sexton, Phyllis E. Singer, and Lars Vilhuber. 2020. "The modernization of statistical disclosure limitation at the U.S. Census Bureau." U.S. Census Bureau Working Paper.

Abowd, John M., Ian M. Schmutte, William N. Sexton, and Lars Vilhuber. 2019. "Why the Economics Profession Must Actively Participate in the Privacy Protection Debate." *AEA Papers and Proceedings*, 109: 397–402.

Benedetto, Gary, Jordan Stanley, and Evan Totty. 2018. "The Creation and Use of SIPP Synthetic Beta v7.0." Center for Economic Studies, U.S. Census Bureau CES Technical Notes Series 18-03.

Biemer, Paul P. 2010. "Total survey error: Design, implementation, and evaluation." *Public Opinion Quarterly*, 74(5): 817–848.

Cohen, Aloni, Moon Duchin, JN Matthews, and Bhushan Suwal. 2021. "Census TopDown: The Impacts of Differential Privacy on Redistricting." 2nd Symposium on Foundations of Responsible Computing.

Ennis, Sharon R., Sonya R. Porter, James M. Noon, and Ellen Zapata. 2018. "When race and Hispanic origin reporting are discrepant across administrative records and third party sources: Exploring methods to assign responses." *Statistical Journal of the IAOS*, 34(2): 179–189.

Garfinkel, Simson L., John M. Abowd, and Sarah Powazek. 2018. "Issues Encountered Deploying Differential Privacy." U.S. Census Bureau Working Paper.

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

Hawala, Sam. 2008. "Producing Partially Synthetic Data to Avoid Disclosure." Proceedings of the Joint Statistical Meetings.

Hotz, V. Joseph, Christopher R. Bollinger, Tatiana Komarova, Charles F. Manski, Robert A. Moffit, Denis Nekipelov, Aaron Sojourner, and Bruce D. Spencer. 2022. "Balancing data privacy and usability in the federal statistical system." *Proceedings of the National Academy of Sciences*, 119(31): e2104906119.

Jarmin, Ron. 2019. "Census Bureau Adopts Cutting Edge Privacy Protections for 2020 Census." *https://www.census.gov/newsroom/blogs/random-samplings/2019/02/census$_b$ureau$_a$dopts.html*, *Accessed* : 2022 − 04 − 21.

Kenny, Christopher T., Shiro Kuriwaki, Cory McCartan, Evan T. R. Rosenman, Tyler Simko, and Kosuke Imai. *2021. "The use of differential privacy for census data and its impact on redistricting: The case of the 2020 U.S. Census."* Science Advances*, 774(41).*

Meyer, Bruce D., and Nikolas Mittag. *2019. "Using Linked Survey and Administrative Data to Better Measure Income: Implications for Poverty, Program Effectiveness, and Holes in the Safety Net."* American Economic Journal: Applied Economics*, 11(2): 176–204.*

Meyer, Bruce D., and Nikolas Mittag. *2021*a. *"Combining Administrative and Survey Data to Improve Income Measurement." In* Administrative Records for Survey Methodology*. , ed. Asaph Young Chun, Michael D. Larsen, Gabriele Durrant and Jerome P. Reiter, Chapter 12, 297–322. John Wiley & Sons.*

Meyer, Bruce D., and Nikolas Mittag. *2021*b. *"An empirical total survey error decomposition using data combination."* Journal of Econometrics*, 224(2): 286–305.*

Meyer, Bruce D., Nikolas Mittag, and Robert M. George. *2022. "Errors in Survey Reporting and Imputation and their Effects on Estimates of Food Stamp Program Participation."* The Journal of Human Resources*, 57(5): 1605–1644.*

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

Ruggles, Steven, Catherine Fitch, Diana Magnuson, and Jonathan Schroeder. *2019. "Differential Privacy and Census Data: Implications for Social and Economic Research."* AEA Papers and Proceedings*, 109: 403–408.*

Santos-Lozada, Alexis R., Jeffrey T. Howard, and Ashton M. Verdery. *2022. "How differential privacy will affect our understanding of health disparities in the United States."* PNAS*, 117(4): 13405–13412.*

Winkler, Richelle L., Jaclyn Butler, Katherine J. Curtis, and David Egan-Robertson. *2022. "Differential Privacy and the Accuracy of County-Level Net Migration Estimates."* Population Research and Policy Review*, 41(41): 417–435.*