Statistical Disclosure Limitation and Total Survey Error*

Evan Totty
U.S. Census Bureau

Thor Watson
U.S. Census Bureau

May 6, 2024

Abstract

The effect of privacy protection on survey data accuracy is not well understood, especially relative to other sources of survey error. A common evaluation approach is to compare statistics generated from the survey with versus without privacy protection. However, this approach implicitly assumes that the survey before privacy protection is the "truth," which we know is not often the case. When there is already error in the survey, differences between the original and privacy-protected data do not directly translate to accuracy reductions. We demonstrate an improvement to this approach by extending the total survey error framework to include error from privacy protection and applying it to linked survey-administrative data with and without synthesis applied to the survey. Doing so allows us to evaluate the relative magnitude of coverage error, measurement error, item non-response error, and privacy protection error. We also evaluate the differential effect of these errors on demographic sub-groups, which impacts frequently used measures of inequality. Using American Community Survey data linked to administrative and proprietary data, we find that error from privacy protection in a select set of estimated means and population sizes is smaller on average than measurement error or coverage error and similar in magnitude to non-response error. Error from privacy protection tends to be larger for categorical variables than continuous variables, while the opposite is true for measurement error and for non-response error. Additionally, error from privacy protection tends to shrink estimated outcome gaps between demographic sub-groups, whereas measurement error tends to inflate those same gaps.

*This paper is being released to inform interested parties of ongoing research and to encourage discussion of work in progress. We thank Gary Benedetto, Michael Freiman, and Jordan Stanley for their feedback on the paper. Any opinions and conclusions expressed herein are those of the authors and do not represent the views of the Census Bureau or other organizations. The Census Bureau has reviewed this data product to ensure appropriate access, use, and disclosure avoidance protection of the confidential source data used to produce this product. This research was performed under Census project P-7530852. Data from the paper are confidential. (Disclosure clearance numbers: CBDRB-FY24-CED010-0001, CBDRB-FY24-CED010-0002, CBDRB-FY24-CED010-0003).

1 Introduction

Federal statistical agencies collect and disseminate survey data on virtually every aspect of the economy and society of the United States. As part of this process, every survey undergoes some type of privacy protection between data collection and microdata dissemination in an attempt to prevent the public from being able to re-identify survey respondents in the data, a process known as statistical disclosure limitation (SDL). Little is known about how SDL impacts data accuracy, especially in comparison to other components of survey sampling and non-sampling error that are more commonly studied.[1] Rising survey non-response rates and measurement error have led researchers to conclude that household surveys are in a time of "crisis" (Meyer, Mok and Sullivan, 2015). They have also led to growing interest in transforming the statistical agencies to rely less on survey data (Jarmin, 2019).

The impact of privacy protections have historically received considerably less attention, partly because of a necessary lack of transparency regarding the details of how legacy SDL methods are applied and partly because of a lack of datasets with and without privacy protection available for comparison. However, interest in the application and impact of SDL has grown in recent years. Advances in computer science combined with the growth of access to individual-level data from other sources in the age of "big data" has increased the risk of reconstruction and re-identification attacks (Abowd, 2016). This rising risk was the impetus for disclosure avoidance modernization efforts implemented by the Census Bureau (Abowd et al., 2020). The efforts included the application of differential privacy to the 2020 Decennial Census (Abowd, 2018; Garfinkel, Abowd and Powazek, 2018; Hawes, 2020). Formally private SDL approaches such as differential privacy provide mathematical privacy guarantees, thus allowing for greater transparency regarding the application of SDL. The modernization efforts also include the expanded use of "synthesis" in existing surveys, in which a subset

[1] See Bound, Brown and Mathiowetz (2001) and Meyer, Mok and Sullivan (2015) for summaries of the research literature regarding the extent and impact of non-SDL-related survey error. See Kennickell and Lane (2006), Alexander, Davern and Stevenson (2010), and Abowd and Schmutte (2015) for early work on the impact of SDL-related survey error.

of observed records and/or variables in a microdata file are replaced with modeled values for the sake of privacy. Like differential privacy, synthesis provides opportunities for greater transparency regarding the impact of SDL, particularly when combined with a validation server (Benedetto, Stanley and Totty, 2018; Bowen et al., 2022; Carr, Wiemers and Moffitt, 2023; Barrientos et al., 2024; Stanley and Totty, 2024). Synthesis also plays a key role in discussions related to tiered data access (Abraham, 2019; Vilhuber, 2020).[2]

Regardless of the specific SDL method, a careful assessment of its impact on data accuracy is crucial. Such assessments are challenging due in part to the lack of a framework that recognizes other sources of survey error already present in the data before applying SDL. When a dataset is available both with and without privacy protection applied, a common practice is to compare statistics calculated on the data with versus without the privacy protection. However, this comparison ignores other sources of survey error already present in the data and implicitly treats the survey data before applying SDL as if it were the "truth" or represented maximum accuracy. As a result, differences between statistics generated on the original versus protected data are often interpreted as changes from zero error to non-zero error. Such differences do not directly correspond to reductions in data accuracy if the data already contain errors, which we know is the case. A holistic approach that quantifies errors due to SDL relative to errors from other sources is therefore crucial for understanding the effect of SDL on data accuracy.

Rather than simply comparing the survey with versus without applying SDL, we propose linking the survey to population-level administrative data and using those data as a proxy for the truth in order to assess the impact of SDL on survey accuracy. To do so, we borrow the

[2]Currently there are two primary data dissemination approaches to providing microdata access used by agencies such as the Census Bureau: public microdata, which is altered for privacy using a variety of legacy SDL techniques, and access to Federal Statistical Research Data Centers (FSRDCs), which provide access to the unaltered internal confidential data but come with high barriers to use in the form of monetary and time costs. Tiered access aims to fill in the gap between these two dissemination approaches, thereby providing more equitable access to high quality data. One example would be a *fully* synthetic public microdata file with a validation service. Users can build their analysis on the public synthetic file, which may include more variables, sample size, and/or detail than a typical public microdata file due to the increased privacy protection offered by full synthesis. Users can then "validate" their results by having their code run on the internal confidential data and receiving results based on the internal data.

total survey error framework from prior research on survey error that quantifies traditional sources of survey error such as coverage error, measurement error, and item non-response error (Biemer, 2010; Groves and Lyberg, 2010; Meyer and Mittag, 2021*b*). We extend the framework from Meyer and Mittag (2021*b*) to include error from SDL. We then apply this framework to linked survey-administrative data with SDL applied in the form of synthesis. Comparing 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 communicating data quality and uncertainty to data users. Quantifying each type of survey error can be challenging given data requirements, but linked survey-administrative data make this possible.

After extending the framework, we apply it to American Community Survey (ACS) data linked to several different administrative or proprietary datasets. The administrative and proprietary datasets provide population-level proxies for the true information that ACS variables are intended to measure and thus provide an estimate of the survey target. We synthesize the ACS variables of interest using classification and regression tree (CART) synthesis methods. Afterwards, we have three different sources of data for a given variable of interest: the administrative data, the survey data, and the synthetic survey data. We then use these data to analyze the effect of coverage error, non-response error, measurement error, and SDL error on simple but important statistics such as variable means (of wage and salary income, retirement income, home value, property taxes, and birth year) and population sizes of demographic groups (based on race categories, Hispanic status, and citizenship status). We also analyze how the various sources of error differentially impact demographic sub-groups and thereby influence important estimates of inequality.

When we apply the extended total survey error framework to these data we find that SDL error ranges from -14.29% to 4.65% of the survey target depending on the variable, with an average percentage error (APE) of -0.72% or an average absolute percentage error (AAPE) of 3.17%. As a result, SDL error is smaller than the impact of coverage error (-1.00% APE,

12.14% AAPE) and measurement error (5.14% APE, 7.64% AAPE). It is also smaller than the impact of non-response error in terms of APE, but non-response error is smaller in terms of AAPE (1.31% APE, 1.67% AAPE). Furthermore, we find some important heterogeneity between the impact of errors from SDL compared to measurement error and non-response error. First, error from SDL in our data tends to be larger for categorical variables than continuous variables, while the opposite is true for measurement error and for non-response error. For measurement error and non-response error, this may reflect relative ease for survey respondents to recall fixed categorical information like race compared to variable continuous information like income and/or a difference in perceived sensitivity between demographic information versus financial information that might cause respondents to misreport information. For SDL error, this observed difference between continuous and categorical variables likely reflects difficulty in accurately synthesizing categorical information such as race relative to continuous information such as income. Second, measurement error tends to inflate estimated differences in outcomes across demographic sub-groups in our data, whereas SDL error tends to shrink estimated differences in outcomes across the sub-groups. Using income as an example, the inflation in differences across groups due to measurement error can be explained by the fact that sub-groups with more income on average in our data also have larger absolute over-reporting of income on average. For SDL error, the shrinking of differences can be explained by the fact that synthesis can often have a "mean-reverting" effect across sub-groups during the synthesis process, which we will discuss in more detail later in the paper.

Our work contributes to a growing literature on the economic analysis of SDL. Bowen et al. (2022), Carr, Wiemers and Moffitt (2023), and Stanley and Totty (2024) compare statistics derived from confidential microdata to the same statistics derived from a synthetic version of the same data. Abowd and Schmutte (2015), Komarova and Nekipelov (2022), and Agarwal and Singh (2024) consider the validity of econometric approaches applied to privacy protected data. 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. None of these papers directly quantify the impact of error due to SDL relative to, or conditional on, other types of survey error. We argue that our 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.[3] For example, agencies often report estimates of sampling error to go along with published statistics. Similarly, microdata users are accustomed to using methods that generate standard errors and confidence intervals that account for sampling error. Non-sampling errors are measured and accounted for less often, but data providers and users are accustomed to accepting the presence of these errors, whether implicitly or explicitly. Manski (2015) renewed the call from Morgenstern (1963) for statistical agencies to measure and communicate non-sampling errors. There has been a growing effort to understand non-sampling errors both inside and outside of statistical agencies, made possible by the growth in data availability and data sharing agreements between organizations. Such agreements make "matched" or "validation" studies possible, in which surveys are linked to population-level data in order to validate the survey responses (Abowd and Stinson, 2013; Bollinger et al., 2019; Klee, Chenevert and Wilkin, 2019; Meyer and Mittag, 2021*a*,*b*; Rothbaum and Bee, 2021; Meyer, Mittag and George, 2022). The same ideas apply to error from SDL, which is another form of non-sampling error (Hotz et al., 2022). The modernization of disclosure avoidance and movement toward transparency provides an opportunity to incorporate disclosure avoidance into non-sampling error evaluations.

Finally, an important caveat of our approach is that the framework implicitly assumes that the population-level administrative data accurately represents the truth. We under-

[3]For example, an agency with a budget constraint may have to choose between spending additional resources on non-response follow-up in order to increase response rates versus spending on questionnaire design in order to increase response accuracy. Weighing these trade-offs in survey design and implementation requires information on the extent and effect of different sources of error.

stand this is an imperfect assumption, given that administrative data can have their own data quality issues and the target population of the survey is not always exactly the same as that of the administrative data. However, we believe this approach still represents an improvement over assuming that the survey data before applying SDL is the truth and then simply comparing statistics generated from the data with versus without SDL. Comparing deviations between survey and administrative data to deviations between original and synthetic survey data provides a more complete perspective of survey error that gives data providers and data users additional benchmarks for determining acceptable levels 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 describes the datasets we use, provides some linkage details, and summarizes the synthesis methodology. Section 4 discusses the results. Section 5 concludes.

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 with and without privacy protection using SDL, 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 and extensions can be found in Meyer and Mittag (2021*b*).

2.1 Total Survey Error Using Combined Data

The survey data are a sample of the population, while the administrative data contain information for the entire population. Let $\mathcal{P}$ be the population that the survey is intended

to represent and let $\mathcal{S}$ be the population that the survey actually represents. $\mathcal{S}$ may differ from $\mathcal{P}$ due to factors such as frame error or unit non-response. The survey contains person final weights, w_i^f, which weight the observations such that they represent $\mathcal{S}$. Let $P = |\mathcal{P}|$ denote the population size realization assumed by the administrative data and let P^S denote the population size realization assumed by the survey, which is equal to the sum of the weights. Additionally, let $\mathcal{L}$ denote the sub-set of survey observations that have a linkage key and therefore can be linked to the administrative data and assume that $\mathcal{L}$ can be re-weighted using a weight adjustment, $\hat{w}_i$, to account for missing linkages so that it is representative of $\mathcal{S}$. 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 $X = [x_1, ..., x_n]$, where n is the number of individuals in the survey and x_i is the value of X for individual i. To fix ideas, consider an application to wage and salary income such that x_i will be an individual's wage and salary income and μ the estimated average wage and salary income for the population. While we focus on average error in means, one could also analyze many other topics and statistics. For example, μ could be a measure of variance or mean squared error.

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 privacy protected survey data with SDL.[4] Only individuals who can be linked,

[4]We assume that there exists a person-level link between the original data and the privacy protected data with SDL. That is, each record in the privacy protected data corresponds to a particular record in the original survey data. This does not have to be the case. When performing full synthesis, the data could be modeled in such a way that a synthetic record does not correspond directly to any underlying original record. This was essentially the original idea for synthesis in Rubin (1993): multiple imputation could be used, in essence, to complete the missing survey responses for the entire population from which the original sample had been drawn. However, most actual applications of synthesis ended up following the approach from Little (1993), which was to replace original, non-missing values in order to "mask" sensitive values. Therefore, synthesis is often like other SDL techniques in that it does have a one-to-one link between the original data and the privacy protected data. Furthermore, there is always a one-to-one link when the data are partially synthetic, which has been a common use of synthetic data to date (Hawala, 2008; Benedetto, Stanley and Totty, 2018). When there is not a one-to-one link between the original and privacy protected survey data, the framework could still be used for some statistics but not others. For example, the framework could still be used for mean error as long as a linkage status variable used for sub-setting the survey is available, but the framework could not be used for mean squared error due to the summation rules that arise in equation (2) below.

$i \in \mathcal{L}$, have a x_i^A measure. Furthermore, let $\hat{\mu}^A$ be the population average for the administrative records and $\hat{\tilde{\mu}}^S$ be the weighted survey estimate on the privacy protected data.[5] We assume that the administrative records are accurate, meaning that there is no error in the administrative records such that any discrepancies between the survey and administrative data are a result of survey error.[6] We are following Meyer and Mittag (2021*b*) in making this assumption, but we acknowledge that assuming administrative records do not contain errors is a flawed assumption as it is well-documented that administrative records 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. We leave work that accounts for administrative records error within this framework for future work.

For a given target parameter μ, the total survey error is the difference between the true population parameter and the survey estimate. We can therefore 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 (privacy protected) survey-weighted estimate of the target parameter. This part is estimated from the linked sample, $\mathcal{L}$, using both the final survey weights and the weight adjustments to account for the missing linkages.[7] The second

[5]The framework assumes that the administrative data are accurate, but we use $\hat{\mu}^A$ rather than μ^A here to indicate that some components of the administrative data mean may have to be estimated. For example, when the administrative data cover individuals or households that the survey was not intended to cover, these groups have to be identified and dropped from the administrative source, which is a process that may induce error.

[6]When an individual from $\mathcal{L}$ appears in the administrative data, discrepancies are assigned to survey error. When an individual appears in $\mathcal{L}$ but not the administrative data, the treatment depends on the variable type. For continuous amounts that only apply to a subset of individuals, such as wage and salary income, we assume it must be a false positive error in the survey data; i.e., the individual does not exist in the administrative data for wage and salary income because they did not earn any wage and salary income. For categorical information that applies to all individuals, such as race or Hispanic status, we cannot infer the explanation and therefore we drop the individual from $\mathcal{L}$.

[7]All summations are over $\mathcal{L}$ unless otherwise indicated.

part of the expression is the population average from the 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 a survey excludes group quarters, then individuals living in group quarters should ideally be removed from the calculation on the administrative data when such individuals can be identified in the administrative data. Furthermore, administrative records that are not linkable (e.g., observations in the data that are missing a linkage key) should be excluded from the administrative data because the survey data cannot cover them. We describe our data cleaning process in further detail in Appendix A.

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 $x_i^A + ((1-r_i)+r_i)(x_i^S - x_i^A) + (\tilde{x}_i^S - x_i^S)$ to produce 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 average when the survey responses are replaced with administrative data values for the linked records. Generalized coverage error represents the combination of frame error and unit non-response error.[8] Generalized coverage error also includes sampling error in the survey estimate of

[8]The final survey weights, which are based on the initial survey weight equal to the inverse probability

μ. The second term is item non-response error ($\hat{\varepsilon}_{INRE}$), which is the weighted average 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 average 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 average difference between the individual's response in the privacy protected and original survey data.

Generalized coverage error, item non-response error, and measurement error are each a function of the difference between the administrative records and the original survey data (rather than the privacy protected survey data) because these error components are each part of the survey design and implementation that occur before applying SDL. SDL error is a function of the difference between the privacy protected 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 mimicking that data and minimizing the impact on statistics derived from the survey subject to disclosure avoidance constraints.

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

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

of selection and adjustments for unit non-response, are intended to prevent coverage error.

where

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

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

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

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 with covariates Z_i.[9] That is, $\mathbb{E}(x_i^S | i \in \mathcal{L}, Z_i) = \mathbb{E}(x_i^S | Z_i)$ and $\mathbb{E}(\tilde{x}_i^S | i \in \mathcal{L}, Z_i) = \mathbb{E}(\tilde{x}_i^S | Z_i)$. Consistency of the estimates for generalized coverage error, measurement error, and item non-response error also require $\mathbb{E}(x_i^A | i \in \mathcal{L}, Z_i) = \mathbb{E}(x_i^A | Z_i)$, $\mathbb{E}(x_i^A | i \in \mathcal{L}, Z_i, r_i = 1) = \mathbb{E}(x_i^A | Z_i, r_i = 1)$, and $\mathbb{E}(x_i^A | i \in \mathcal{L}, Z_i, r_i = 0) = \mathbb{E}(x_i^A | Z_i, r_i = 0)$, respectively. Alternatively, the survey mean and SDL error could be estimated based on the full survey sample rather than the linked sample. The advantage of using the linked sample is that it holds the sample constant across all the different survey error calculations so that the relative error amounts are directly comparable and also sum to the total survey error [i.e., the components from equation (2) sum to the total from equation (1)].

[9] We adjust the weights using inverse probability weighting. More specifically, we establish a list of covariates, Z_i, that we think may be correlated with whether an individual in the survey is assigned a linkage key (LINK) and then regress LINK status on these covariates for all individuals. The inverse of the predicted linkage probability from this regression, $1/\Pr(LINK = 1)$, is our weight adjustment, $\hat{w}_i$. We use age, age-squared, race, sex, Hispanic status, marital status, citizenship status, education level, number of persons in the household, poverty status of the household, and state of residence as the covariates in Z. For analysis of household-level variables (home value and property tax), no linkage adjustment is necessary because the home addresses randomly sampled for the survey in the first place are derived from a master address file. Consequently, all households in the survey are assigned a linkage key that is tied to their mailing address. More information on the linkage keys is provided in section 3.1.

3 Data Sources, Linkage, and Synthesis

3.1 Data Sources and Linkage

In this paper, we analyze the largest ongoing household survey conducted by the U.S. Census Bureau: the American Community Survey (ACS). The ACS randomly samples approximately 295,000 household addresses each month across the United States with no address being selected more than once every five years (U.S. Census Bureau, 2014). The survey is primarily conducted through an online portal or a paper form mailed out to target households, but in some situations a telephone or in-person interview may be performed (e.g., group housing addresses or submitted questionnaires that require some clarification). The survey questionnaire is relatively long and includes detailed questions concerning demographic and housing characteristics. For the task at hand, we only use data from the 2019 ACS but plan to expand our analysis to include additional survey years in the future.

The survey variables we are interested in examining from the ACS include wage and salary income, retirement income, home value, property tax, year of birth, race and ethnicity, and citizenship status.[10] For each variable, however, we also require a measure of the truth at the individual or household level for the entire U.S. population in order for us to analyze the total survey error. To that end, our population data are derived from various administrative and proprietary data sources. More specifically, we use W-2 and 1099-R forms from the Internal Revenue Service (IRS) to measure wage and retirement income; we use a proprietary valuation model from Black Knight, Inc. to measure home value; we use public tax records collected by Black Knight, Inc. to measure property taxes[11]; we use records from the Social Security Administration (SSA) to measure year of birth and citizenship status; and lastly, we

[10] See Table A1 for the original phrasing of each survey question from which these variables are derived.

[11] The Black Knight, Inc. property tax data is constructed partially from property assessments and partially from tax bills. That is, some of the tax amounts are obtained from assessed property taxes (before any exemptions or appeals) while other tax amounts are from the actual tax bill (after any exemptions or appeals). The latter aligns much better with the survey question, so we base our tax analysis on the subset of Black Knight, Inc. data derived from the tax bills. In doing so, we are assuming that the property tax source is as good as random. The source is largely dependent on the state and county due to natural variation in the date of billing.

use the Census Bureau's Best Race and Ethnicity internal file to measure race and ethnicity.

We link the ACS to the administrative data at the individual level using unique person identifiers created by the Census Bureau's Person Identification Validation System (PVS). The PVS was developed in 1999 as a collaboration between the Census Bureau and the Social Security Administration. It uses probabilistic linking to match person-level survey data to a reference file that contains one record for each Social Security number while keeping all variations of an individual's name, date of birth, and address information in separate files (Wagner and Layne, 2014). A matched person record is assigned a unique person identifier called a protected identification key (PIK) that is akin to an anonymized Social Security number. A PIK can be used as a linkage key across all files that have ever been processed by the Census Bureau using the PVS. In our context, the PIK rate of the ACS person-level file is 91.98% while the IRS W-2 forms, IRS 1099-R forms, SSA records, and the Best Race and Ethnicity internal file have PIK rates of 99.99%, 100%, 100%, and 100% respectively.

For the housing variables, we link the ACS to the Black Knight, Inc. data at the housing level using unique domicile identifiers created by the Census Bureau's Master Address File (MAF). The MAF was initially created by the Census Bureau, in collaboration with the U.S. Postal Service and local governments, to improve the mailing list for the 2000 Decennial Census. Ultimately, it became a complete repository of every residential, and select non-residential, mailing address in the United States that is updated on a semiannual basis (National Academies of Sciences and Medicine, 2023). Every domicile in the MAF is assigned a unique identifier, called the MAFID, that can be used as a linkage key across household surveys and records processed by the Census Bureau. In our context, a MAFID is assigned to 100% of records in the ACS housing-level file while the Black Knight, Inc. housing data have a MAFID assigned 63.33% of the time.[12] See Appendix A for additional details on our data sources and linking process.

[12]The Black Knight, Inc. housing data have a relatively low MAFID match rate because the raw data include both commercial and residential properties, but the MAF is primarily focused on capturing the universe of residential mailing addresses and only includes some non-residential addresses.

3.2 Synthesis

We created the synthetic version of the survey variables in Table A1 using the Census Bureau's data synthesizer, known as CenSyn. CenSyn uses CART methods to replace observed survey values with modeled values. Synthesis can be either "partial synthesis," meaning that only a subset of variables and/or observations are synthesized, or "full synthesis," meaning that all variables and observations are synthesized.[13] There is precedent at the Census Bureau for both types of synthesis. Partial synthesis is more commonly applied within the typical production process and is considered to be one of many legacy SDL techniques available for privacy protection (U.S. Census Bureau, 2019). Full synthesis is currently less common and usually involves a validation server that allows users to obtain results based on the non-synthetic internal data by sending their code to the Census Bureau after developing it on the synthetic data. For this reason, when analyzing a given variable of interest we synthesize the variable as if it were being created for a partially synthetic dataset in which only that variable was synthetic.[14] This point is relevant when we evaluate the differential impact of synthesis for a variable of interest, Y, on sub-groups defined by variable G because it means G is non-synthetic even if it is used as the synthetic variable of interest elsewhere in the analysis. A benefit of this approach is that it avoids issues related to the order of synthesis when multiple variables are synthesized together, which can sometimes influence the relative quality of different synthetic variables.

For a given variable that is being synthesized, CenSyn builds trees to predict the value of that variable based on available covariates (also known as dependencies). CenSyn uses CART to recursively partition the data into nodes in order to maximize homogeneity of

[13] See Drechsler and Haensch (2023) for more details on the history, methodologies, and uses of synthetic data.

[14] There are two exceptions: one for home value and property taxes and another for race and Hispanic status. Home value and property tax are related by definition, so we synthesize them as such by first synthesizing home value and then synthesizing property tax based on synthetic home value. Race and Hispanic status are synthesized in a similar way, in which we first synthesize race and then synthesize Hispanic status based on synthetic race. We do this because we combine race and Hispanic status when evaluating the data later in the paper, so this relationship also needs to be modeled in synthesis.

the variable being synthesized within the node. This process continues and the trees grow deeper until some stopping criterion is reached, at which point the nodes become terminal nodes, also known as "leaves." After the trees are fit on the original data, each observation in the dataset is sent through the tree to a leaf based on that observation's values for the dependencies that appear in the trees. The observation's *original* values for the dependencies are used when sending the observation through the tree unless multiple variables are being synthesized in a sequential and dependent process. In that case, the *synthetic* values for the dependencies are used for any previously-synthesized variables that appear in the trees. When a leaf is reached, a value from the original data is drawn at random, with replacement, from within the leaf. This value is the new "synthetic value" for that record and replaces the original value of the variable in the dataset.

Several possible stopping criteria are available to CenSyn, including maximum tree depth, minimum leaf size, minimum node size for further splits, and minimum threshold for homogeneity improvement (Breiman et al., 1984). These criteria essentially serve as the privacy parameters when creating synthetic data with CenSyn: they limit the ability of the trees to perfectly partition the data into leaves with complete homogeneity, which would reproduce the original value with certainty for records that fall into such leaves. Shallower trees and leaves containing a greater number of observations will tend to introduce larger amounts of "noise" or "uncertainty" into the data by generating more heterogeneity within the leaves and in turn allowing for more possible values that can be drawn from a given leaf. The accuracy of the data is therefore influenced by the privacy parameters as there is in an inherent trade-off between introducing more noise into the data and reproducing original records.

Separate from the privacy parameters, accuracy of the data is also influenced by the covariates available to CenSyn as dependencies or "predictors" when building the trees. CenSyn requires a user to specify the set of possible covariates available as predictors. An important aspect of the CART-based approach to synthesis used by CenSyn is that covariate relationships from the original data will be maintained in the synthetic data only to the extent

that they involve covariates that appear in the trees (or to the extent that they are correlated with other covariates that appear in the trees). Therefore, covariate relationships that are related to high priority use cases of the data need to be accounted for when specifying the set of covariates available for building the tree.

Table A2 summarizes the basic details and dependencies for each variable that was synthesized. It includes the variable type, such as categorical or numerical; the possible variable values, which determine the range of possible synthetic values; and the set of dependencies used for building the trees.[15] In addition to the details listed in Table A2, each variable was modeled with state of residence as an "independent feature," meaning that separate trees were trained independently for each state. We imposed relatively few constraints on the trees via the privacy parameters, requiring only that there be at least five records in each leaf. This ensures that each observation receives a synthetic value that is a random draw from at least five different original values. This constraint alone is enough to guarantee some uncertainty has been introduced into the data regarding the true value of that variable from the perspective of a data user who has access to the microdata file with the synthesized value.

Another important aspect of CART-based synthesis is that covariates that are more useful for partitioning the data into homogeneous nodes will be prioritized by CenSyn when building the trees, but those covariates may not always be the same covariates as those from high priority use cases. For this reason, it can sometimes be useful to pare down the list of predictors to those most related to high priority use cases in an attempt to "force" CenSyn to split on those predictors and thus hopefully maintain covariate relationships in the synthetic data that are most related to high priority use cases. However, it is not realistic to always force the trees to perfectly partition the data based on a particular covariate even with a limited list of dependencies. If all the leaves are not perfectly partitioned with respect to a

[15]For variables with end points that contain special mass in the distribution, such as $0 and the top-code value of $999,999 for wage and salary income, CenSyn can first model whether a record belongs to either of the end points and then model the rest of the distribution. We used the end points modeling for the top- and bottom-code values for wage and salary income, retirement income, home value, and property tax.

particular covariate, then synthesis will often attenuate the relationship between the variable being synthesized and the given covariate to some degree. For instance, suppose we use sex as a predictor when synthesizing income. If all leaves are perfectly partitioned based on sex, then sex is perfectly accounted for when predicting/synthesizing income, and sex-based differences in the distribution of income should be fully reproduced in the synthetic data (aside from differences that arise due to randomness in the draws from the leaves). On the other hand, if some leaves are not partitioned based on sex, yet males still have a different distribution of income in that leaf compared to females, then males and females in that leaf will receive draws of synthetic values based on a combination of the two distributions. In this scenario, male-female income differences for records in those particular leaves would be removed in the synthetic data, thus attenuating the relationship in the dataset as a whole. We refer to this aspect of CenSyn as the "mean-reverting" nature of synthesis.

Clearly, the accuracy of synthetic data depends critically on the quality of the models used to generate the synthetic data. We acknowledge that our analysis below is arguably a relatively simple test of the models in that we evaluate few use cases compared to the full set of possible use cases of ACS data and that they are based on relatively simple statistics such as means and counts. However, the fact that we test relatively few use cases is due to a lack of population-level administrative data currently available for other variables rather than a limitation of the models. The focus on relatively simple statistics is due to the fact that it keeps the total survey error framework tractable and the fact that other sources of non-sampling error have been shown to bias even simple statistics (Meyer and Mittag, 2019, 2021*a*,*b*). Stanley and Totty (2024) assess the similarity between synthetic survey data and its non-synthetic counterpart for a larger number of use cases and more complex statistics, although they do not account for other sources of error already in the survey.

Finally, we create multiple implicates of each synthetic variable by running the synthesizer five separate times to create five separate synthetic versions of the variable. All results for synthetic variables in the paper are based on computing the given statistic on each implicate

and then averaging across all five implicates.

4 Results

4.1 Total Survey Error

Table 1 and Table 2 summarize the total survey error and its four components for eleven different statistics. Table 1 shows results for the numeric variables. Table 2 shows results for the categorical variables, with race and Hispanic status combined into five different groups: White non-Hispanic, Black non-Hispanic, Asian non-Hispanic, other race non-Hispanic, and Hispanic. In each table, panel A shows the variable mean in the administrative dataset, the survey, and the synthesized version of the survey.[16] Meanwhile, panel B shows the traditional components of survey error, based on comparing the survey dataset to the corresponding administrative dataset that serves as our proxy for the true survey target, and the "total" survey error before the application of SDL in the form of synthesis. Lastly, panel C shows the extended survey error component for SDL, which is based on comparing the original survey records to the synthetic survey records, and the total survey error after the variable is synthesized. Error amounts are shown first in real amounts (dollars, years, or percentage points, depending on the variable being analyzed) and then also shown in relative terms as a percentage of the variable mean in the administrative dataset (i.e., the survey target). The percentage error amounts are reported in parentheses underneath the real amounts.

Focusing first on wage and salary income in Table 1, panel A shows that average wage and salary income is less in the administrative data than in the survey data ($28,760 ver-

[16]Note that for wage and salary income and retirement income the results report mean income for the entire in-scope population in the survey, which is individuals age 15 and older. Thus, these are means for the whole target population (inclusive of zeros) rather than just for individuals with positive income amounts. Similarly, the means for home value and property tax are means for the entire survey target population, which is owner-occupied homes. Thus, for all four variables, the means are based on summing the wage/retirement/tax/value amounts in the given source (unweighted amounts in the population source and weighted amounts in the survey source) and then dividing by the size of the target population, which is equal to the sum of the survey weights for the in-scope observations. The remaining variables are in-scope for all observations, so the means are simply the variable mean in each data source.

sus $33,220) and synthesis further increases the survey mean by a relatively small amount ($33,220 versus $33,430). Panel B shows that coverage error is -$1,233, meaning that individuals who ended up in the survey sample have less income on average than the population as a whole according to the administrative dataset; measurement error is $4,237, meaning that individuals who responded to the survey over-report their income on average relative to the amounts found in the administrative data; and non-response error is $1,450, meaning that imputed income amounts for individuals who did not respond to the wage and salary survey question were larger than the actual amounts found in the administrative data on average. Totaling these three error components implies that the survey estimate of average wage and salary income is over-estimated by $4,464 before synthesis is applied. Finally, panel C shows that SDL error is $207.10, meaning that individuals in the survey have more income on average after synthesis than before synthesis. This raises the total survey error to $4,671.10.

Moving beyond wage and salary income, there are several points to emphasize based on the entire set of results in Table 1 and Table 2. First, regarding the traditional error components, non-response error is often the smallest in absolute terms (ten out of eleven total statistics), while coverage error is often the largest (seven out of eleven total statistics). Second, SDL error is often among the smallest of all the error components. In particular, SDL error is the smallest for four of the statistics across both tables: mean wage and salary income, mean home value, mean property taxes, and proportion citizen; meanwhile, SDL error is the largest error component only twice: proportion Black and proportion Hispanic. Third, SDL error can sometimes offset a portion of the error from the other components, such that the total survey error is reduced after synthesis. This is the case for mean property taxes, proportion White, and proportion Other Race. The fact that total survey error can be reduced after accounting for SDL error is a clear derivation from the total survey error framework, but it represents a crucial departure from previous work in which any deviation between the original and privacy protected data is implicitly treated as an increase from zero

Table 1: Total Survey Error for Numeric Variables

	(1) Wage & Salary	(2) Retirement Inc.	(3) Home Value	(4) Prop. Taxes	(5) Birth Year
Panel A: Variable means from different sources					
Adrec Mean	$28,760	$3,407	$310,200	$2,435	1983
Survey Mean	$33,220	$3,349	$332,600	$3,040	1982
Synthetic Mean	$33,430	$3,499	$333,600	$3,035	1981
Panel B: Traditional sources of survey error and total					
Coverage Error	-$1,223	-$164.8	-$39,200	-$128.60	-1.195
(% of Adrec Mean)	(-4.25%)	(-4.84%)	(-12.64%)	(-5.28%)	(-0.06%)
Measurement Error	$4,237	-$29.53	$50,460	$617.10	-0.039
(% of Adrec Mean)	(14.73%)	(-0.87%)	(16.27%)	(25.34%)	(-0.00%)
Nonresponse Error	$1,450	$136.2	$11,110	$116.30	0.01
(% of Adrec Mean)	(5.04%)	(4.00%)	(3.58%)	(4.78%)	(0.00%)
Total Error (w/o SDL)	$4,464	-$58.13	$22,370	$604.80	-1.224
(% of Adrec Mean)	(15.52%)	(-1.71%)	(7.21%)	(24.84%)	(-0.06%)
Panel C: SDL error and new total					
SDL Error	$207.1	$149.9	$1,060	-$5.23	-0.992
(% of Adrec Mean)	(0.72%)	(4.40%)	(0.34%)	(-0.22%)	(-0.05%)
Total Error (w/ SDL)	$4,671.1	$91.77	$23,430	$599.57	-2.216
(% of Adrec Mean)	(16.24%)	(2.69%)	(7.55%)	(24.62%)	(-0.11%)

Source: 2019 American Community Survey (ACS), Internal Revenue Service (IRS) W-2 and 1099-R forms for tax year 2018, Black Knight, Inc. housing data, Social Security Administration (SSA) records.

Note: The Adrec Mean is calculated using the entire universe of records from an administrative data source, which can vary across columns, e.g., the administrative data source for Wage & Salary is IRS W-2 forms while SSA records are used for Birth Year. More information on the data can be found in section 3.1. The various error components are defined in section 2.2. The amounts in parentheses are the nominal error amounts converted to percentage error by dividing by the adrec means in Panel A. The Statistical Disclosure Limitation (SDL) method being applied here is synthesis applied to the ACS variable listed in the column heading using classification and regression tree methods to replace observed survey values with modeled values. All results have been reviewed by the U.S. Census Bureau to ensure no confidential information has been disclosed: CBDRB-FY24-CED010-0001, CBDRB-FY24-CED010-0002.

Table 2: Total Survey Error for Categorical Variables

	(1) White	(2) Black	(3) Asian	(4) Other	(5) Hispanic	(6) Citizen
Panel A: Variable means from different sources						
Adrec Mean	0.653	0.126	0.043	0.031	0.179	0.868
Survey Mean	0.611	0.118	0.057	0.035	0.178	0.934
Synthetic Mean	0.638	0.100	0.060	0.035	0.167	0.936
Panel B: Traditional sources of survey error and total						
Coverage Error	-0.0510	-0.0080	0.0140	0.0080	0.0070	0.0390
(% of Adrec Mean)	(-7.81%)	(-6.35%)	(32.56%)	(25.81%)	(3.92%)	(4.49%)
Measurement Error	0.0090	0.0010	0.0010	-0.0030	-0.0080	0.0250
(% of Adrec Mean)	(1.38%)	(0.79%)	(2.33%)	(-9.68%)	(-4.47%)	(2.88%)
Nonresponse Error	0.0010	-0.0005	-0.0001	-0.0002	0.0000	0.0030
(% of Adrec Mean)	(0.15%)	(-0.40%)	(-0.23%)	(-0.65%)	(0.00%)	(0.35%)
Total Error (w/o SDL)	-0.0410	-0.0075	0.0149	0.0048	-0.0010	0.0670
(% of Adrec Mean)	(-6.28%)	(-5.95%)	(34.65%)	(15.48%)	(-0.56%)	(7.72%)
Panel C: SDL error and new total						
SDL Error	0.0270	-0.0180	0.0020	-0.0010	-0.0100	0.0020
(% of Adrec Mean)	(4.13%)	(-14.29%)	(4.65%)	(-3.23%)	(-5.59%)	(0.23%)
Total Error (w/ SDL)	-0.0140	-0.0255	0.0169	0.0038	-0.0110	0.0690
(% of Adrec Mean)	(-2.14%)	(-20.24%)	(39.30%)	(12.26%)	(-6.14%)	(7.95%)

Source: 2019 American Community Survey (ACS), U.S. Census Bureau's Best Race and Ethnicity internal file, Social Security Administration (SSA) records.

Note: The Adrec Mean is calculated using the entire universe of records from an administrative data source, which can vary across columns, e.g., the administrative data source for the race variables is the Best Race and Ethnicity internal file while SSA records are used for citizenship status. More information on the data can be found in section 3.1. The various error components are defined in section 2.2. The amounts in parentheses are the nominal error amounts converted to percentage error by dividing by the adrec means in Panel A. The Statistical Disclosure Limitation (SDL) method being applied here is synthesis applied to the ACS variable listed in the column heading using classification and regression tree methods to replace observed survey values with modeled values. All results have been reviewed by the U.S. Census Bureau to ensure no confidential information has been disclosed: CBDRB-FY24-CED010-0001.

error to non-zero error.

Table 3 reports the average of the percentage error amounts from Table 1 and Table 2, separately for each table and combined across tables. Percentage error allows us to express the error amounts across different variables in comparable terms, which in turn allows us to report averages of the different sources of error across variables. Panel A reports the average percentage error (APE). Panel B reports the average absolute percentage error (AAPE), which can provide a useful distinction when, e.g., a given error type is always large but the direction of the error differs across variables such that the APE appears small. The averages are reported separately for numeric versus categorical variables and combined across all variables.

For numeric variables, SDL error is the smallest error component on average based on both average percentage error and average absolute percentage error (0.91% APE, 0.99% AAPE), followed by non-response error (2.75% APE, 3.05% AAPE), coverage error (-10.78% APE, 10.78% AAPE), and measurement error (11.41% APE, 11.70% AAPE). For categorical variables, non-response error is the smallest (-0.13% APE, 0.30% AAPE), followed by measurement error (-1.13% APE, 3.59% AAPE), SDL error (-2.35% APE, 5.35% AAPE), and coverage error (8.77% APE, 13.49% AAPE). Combined across all variables, SDL error is the smallest error component in terms of APE and the second smallest in terms of AAPE (-0.72% APE, 3.17% AAPE) while non-response error is the smallest in terms of AAPE (1.31% APE, 1.67% AAPE). Measurement error is the largest error component in terms of APE (5.14% APE, 7.64% AAPE) while coverage error is the largest in terms of AAPE (-1.00% APE, 12.14% AAPE).

There are two key takeaways from Table 3. First, SDL error is quite small on average and is generally among the smallest error components across all variables. Second, SDL error is smaller for numeric variables than categorical variables, while measurement error and non-response error are smaller for categorical variables than numeric variables. Larger SDL error for categorical variables likely reflects difficulty in accurately synthesizing demographic

Table 3: Average Error Across Variables

	(1) Numeric	(2) Categorical	(3) Combined
Panel A: Average Percentage Error			
Coverage Error	-10.78%	8.77%	-1.00%
Measurement Error	11.41%	-1.13%	5.14%
Nonresponse Error	2.75%	-0.13%	1.31%
Total Error (w/o SDL)	3.37%	7.51%	5.44%
SDL Error	0.91%	-2.35%	-0.72%
Total Error (w/ SDL)	4.28%	5.16%	4.72%
Panel B: Average Absolute Percentage Error			
Coverage Error	10.78%	13.49%	12.14%
Measurement Error	11.70%	3.59%	7.64%
Nonresponse Error	3.05%	0.30%	1.67%
Total Error (w/o SDL)	12.49%	11.77%	12.13%
SDL Error	0.99%	5.35%	3.17%
Total Error (w/ SDL)	12.76%	14.67%	13.72%

Note: This table reports the average of the percentage error and absolute percentage error amounts from Table 1 (numeric variables), Table 2 (categorical variables), and combined across all variables. All results have been reviewed by the U.S. Census Bureau to ensure no confidential information has been disclosed: CBDRB-FY24-CED010-0001, CBDRB-FY24-CED010-0002.

information such as race relative to financial information such as income. Smaller measurement error and non-response error for categorical variables may reflect the relative ease for survey respondents to recall fixed information like race compared to variable information like income and/or a difference in perceived sensitivity between demographic information versus financial information that might cause users to misreport information.

Finally, tables B1 and B2 in appendix B further break down the results by errors due to false positive self-reports in the survey, false negative self-reports in the survey, and differences in continuous amounts. For example, much of the measurement error and non-response error in wage and salary income is due to false positives in the survey, whereas SDL error is more evenly spread between false positives and false negatives.

4.2 Subgroup Heterogeneity and Effects on Estimated Inequality

In addition to calculating the total survey error for each variable, we also computed the error components separately for different sub-groups based on sex, race, and education level. These groups relate to important measures of inequality, so any differential effects of survey error by sub-group can impact commonly used measures of inequality. Unfortunately, our administrative data sources typically lack demographic information on sex, race, and education, so we cannot compute the administrative variable mean or the generalized coverage error for specific sub-groups (both of which require population-level group-specific means from the administrative data). Moreover, we are unable to compute *total* survey error results comparable to those in the prior section.[17] Instead, we focus on measurement error, non-response error, and SDL error, all of which we can compute without knowing demographic information in the administrative data because these components are only based on the survey sample for which we have demographic information from the survey.

Figures C1 through C28 in Appendix C display the survey error components separately

[17]For example, IRS W-2 forms do not include information on race, sex, or education. We therefore cannot compute race-, sex-, or education-specific results for mean wage and salary income. In the future, we plan to link different sets of administrative data together to potentially circumvent this limitation, such as linking the Census Bureau's Best Race and Ethnicity internal file to the W-2 forms.

for each variable and each sub-group. Given the large number of variable and sub-group combinations, we focus on a subset of key sub-group comparisons in the main text: wage and salary income gaps (also known as "wage gaps") and home value gaps. Table 4 shows the effect of survey error on the male-female wage gap, White-Black wage gap, and college-high school wage gap. Meanwhile, Table 5 shows the same gaps, except for home value rather than wage and salary income. For each of the tables, panel A shows the variable mean in the survey and the synthesized version of the survey for each sub-group. Panel B shows the measurement error, non-response error, and SDL error for each sub-group. Panel C shows the resulting respective gaps in the original version of the survey (based on panel A) and the contribution of measurement error and non-response error to each gap (based on panel B). The contribution of a given error component to the gap is based on subtracting the respective female/Black/high school amounts from the male/White/college amounts in panel B. For example, consider the contribution of measurement error to the male-female wage gap Table 4. Based on the measurement error amount in column (2) of panel B, which shows males over-report their wage and salary income by \$5,318 on average, and the amount in column (1) of panel B, which shows that females over-report their income by \$3,227 on average, panel C shows that measurement error increases the estimated male-female wage gap by \$2,091 (\$5,318 minus \$3,227). Finally, panel D shows the respective gaps in the synthetic version of the survey (based on panel A) and the contribution of SDL error to each gap, based on subtracting the respective female/Black/high school SDLE amount from the male/White/college SDLE amount in panel B.

Focusing first on wage and salary income in Table 4, columns (1)-(2) of panel A show mean income for females and males in the original survey data and synthetic survey data. Females have a mean income of \$24,990 in the original survey data (\$25,270 synthetic) compared to \$42,030 for males (\$42,160 synthetic). Panel B shows the contribution of measurement error, non-response error, and SDL error to those amounts. Measurement error is the largest error component for all six demographic groups in the table. The measurement error and

non-response error contributions are also all positive for all six groups, meaning that the survey values are larger than the administrative values, on average, for both observed and imputed values. SDL error is positive for some sub-groups and negative for others. Column (2) of panel C shows the male-female wage gap based on the original survey data ($17,040) and the contribution of measurement error and non-response error to the gap. Measurement error and non-response error both increase the male-female wage gap. In total, $2,605 of the estimated $17,040 male-female wage gap in the original survey data (15.29%) is an over-estimate due to measurement error and non-response error. Finally, panel D shows the male-female wage gap based on the synthetic survey data ($16,890) and the contribution of SDL error to that gap. SDL error decreases the male-female wage gap, reducing it by $149.80. This reduction offsets some of the positive error in the wage gap due to measurement error and non-response error, thus reducing the total error in the estimated gap from $2,605 to $2,455.20.

Moving beyond wage and salary income, there are several points to emphasize based on the entire set of results in Table 4 and Table 5. First, measurement error and non-response error are positive for all 12 sub-group-by-variable combinations across the two tables, indicating that survey responses are consistently larger than administrative values, on average, across both outcomes and all six demographic groups. Second, measurement error differences across demographic sub-groups increase all six of the estimated gaps between the two tables, whereas non-response error increases the gap in some cases and decreases the gap in other cases. Third, the net effect of measurement error and non-response error is a net increase for all six estimated gaps between the two tables, meaning that each gap is over-estimated due to traditional sources of non-sampling survey error. Fourth, SDL error decreases the estimated gap in five of the six cases. SDL therefore offsets some of the traditional survey error and ultimately reduces the total error in four of the six cases.

The reason that measurement error tends to increase estimated gaps can be seen in panel B of Table 4 and Table 5. Measurement error in wage and salary income and home value is

Table 4: Wage and Salary Income Gaps

	(1)	(2)	(3)	(4)	(5)	(6)
	Sex		Race		Education	
	Female	Male	Black	White	High School	College
Panel A: Mean wage & salary income by group						
Survey Mean	$24,990	$42,030	$24,890	$35,810	$20,900	$52,220
Synthetic Mean	$25,270	$42,160	$26,970	$35,310	$20,840	$52,640
Panel B: Wage & salary income error components by group						
Measurement Error	$3,227	$5,318	$2,965	$4,585	$2,665	$6,792
(% of Survey Mean)	(12.91%)	(12.65%)	(11.91%)	(12.80%)	(12.75%)	(13.01%)
Nonresponse Error	$1,202	$1,716	$2,067	$1,149	$1,545	$2,106
(% of Survey Mean)	(4.81%)	(4.09%)	(8.31%)	(3.21%)	(7.39%)	(4.03%)
SDL Error	$279.5	$129.7	$2,079	-$493.6	-$63.18	$419.2
(% of Survey Mean)	(1.19%)	(0.31%)	(8.35%)	(-1.38%)	(-0.30%)	(0.80%)
Panel C: Original gap and error contribution to gap from traditional sources						
Survey Wage Gap		$17,040		$10,920		$31,320
ME Contribution		$2,091		$1,620		$4,127
(% of Survey Wage Gap)		(12.27%)		(14.84%)		(13.18%)
NRE Contribution		$514		-$918		$561
(% of Survey Wage Gap)		(3.02%)		(-8.41%)		(1.79%)
Total Error (w/o SDL)		$2,605		$702		$4,688
(% of Survey Wage Gap)		(15.29%)		(6.43%)		(14.97%)
Panel D: Synthetic gap and error contribution to gap from SDL						
Synthetic Wage Gap		$16,890		$8,340		$31,800
SDLE Contribution		-$149.8		-$2,572.6		$482.38
(% of Survey Wage Gap)		(-0.88%)		(-23.56%)		(1.54%)
Total Error (w/ SDL)		$2,455.2		-$1,870.6		$5,170.38
(% of Survey Wage Gap)		(14.41%)		(-17.13%)		(16.50%)

Source: 2019 American Community Survey (ACS) and Internal Revenue Service (IRS) W-2 forms for tax year 2018.

Note: The demographic information is derived from the original survey responses; hence, we cannot calculate generalized coverage errors. The Statistical Disclosure Limitation (SDL) method being applied here is synthesis applied to wage and salary income using classification and regression tree methods to replace observed survey values with modeled values. All results have been reviewed by the U.S. Census Bureau to ensure no confidential information has been disclosed: CBDRB-FY24-CED010-0001.

Table 5: Home Valuation Gaps

	(1)	(2)	(3)	(4)	(5)	(6)
	Sex		Race		Education	
	Female	Male	Black	White	High School	College
Panel A: Mean home value by group						
Survey Mean	$316,000	$347,200	$235,500	$333,000	$226,000	$410,800
Synthetic Mean	$317,400	$348,100	$263,400	$329,200	$233,800	$402,400
Panel B: Home value error components by group						
Measurement Error	$45,980	$53,570	$30,410	$51,780	$31,530	$62,250
(% of Survey Mean)	(14.55%)	(15.43%)	(12.91%)	(15.55%)	(13.95%)	(15.15%)
Nonresponse Error	$11,160	$8,838	$11,960	$8,335	$13,260	$7,030
(% of Survey Mean)	(3.53%)	(2.55%)	(5.08%)	(2.50%)	(5.87%)	(1.71%)
SDL Error	$1,439	$977.9	$27,920	-$3,774	$7,795	-$8,411
(% of Survey Mean)	(0.46%)	(0.28%)	(11.86%)	(-1.13%)	(3.45%)	(-2.05%)
Panel C: Original gap and error contribution to gap from traditional sources						
Survey Valuation Gap		$31,200		$97,500		$184,800
ME Contribution		$7,590		$21,370		$30,720
(% of Survey Val. Gap)		(24.33%)		(21.92%)		(16.62%)
NRE Contribution		-$2,322		-$3,625		-$6,230
(% of Survey Val. Gap)		(-7.44%)		(-3.72%)		(-3.37%)
Total Error (w/o SDL)		$5,268		$17,745		$24,490
(% of Survey Val. Gap)		(16.89%)		(18.20%)		(13.25%)
Panel D: Synthetic gap and error contribution to gap from SDL						
Synthetic Valuation Gap		$30,700		$65,800		$168,600
SDLE Contribution		-$461.10		-$31,694		-$16,206
(% of Survey Val. Gap)		(-1.48%)		(-32.51%)		(-8.77%)
Total Error (w/ SDL)		$4,806.9		-$13,949		$8,284
(% of Survey Val. Gap)		(15.40%)		(-14.30%)		(4.48%)

Source: 2019 American Community Survey (ACS) and Black Knight, Inc. home valuation data.
Note: The demographic information is derived from the original survey responses; hence, we cannot calculate generalized coverage errors. The Statistical Disclosure Limitation (SDL) method being applied here is synthesis applied to home value using classification and regression tree methods to replace observed survey values with modeled values. All results have been reviewed by the U.S. Census Bureau to ensure no confidential information has been disclosed: CBDRB-FY24-CED010-0001.

positive for all sub-groups, indicating that individuals over-report amounts in the survey on average. However, the amount of the over-reporting is positively correlated with the quantity of interest, meaning that sub-groups with larger average income/home values also tend to over-report their income/home values by larger amounts.

The reason that SDL error tends to decrease estimated gaps can also be seen in panel B of Table 4 and Table 5. In three of the six cases, SDL error is negative for the sub-group with larger average income/home values but positive for the sub-group with smaller average amounts. In another two cases, SDL error is positive for both sub-groups, but larger for the sub-group with smaller average income/home value. This pattern is what we referred to as the "mean-reverting" nature of CART-based synthesis in section 3.2: when the trees are not perfectly partitioned based on a covariate of interest (such as sex, race, or education), then the synthesis process will tend to attenuate the relationship between the synthetic variable and that covariate to some degree.

5 Conclusion

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 the accuracy of the data. This requires a holistic approach to survey error that moves beyond simply comparing the data with versus without privacy protection. Instead, we must understand the impact of SDL protection relative to, and conditional on, other types of survey error already present in the data. Our work provides an important step in this direction by quantifying error from SDL and comparing it to other sources of non-sampling error using linked survey-administrative data and the total survey error framework.

Our results based on synthesis applied to a select set of variables from the ACS suggest that error from SDL has a smaller average impact on the accuracy of these variables than the impact of coverage error or measurement error and a similar impact to non-response error. Additionally, SDL error sometimes offsets other sources of error and reduces total survey error. Our demonstration that total survey error can be increased *or decreased* after accounting for SDL error represents a crucial departure from the common SDL evaluation approach of only comparing survey statistics generated before versus after applying SDL, in which case any deviation is often interpreted as an increase from zero error to non-zero error. We also demonstrate some important differences between the impact of SDL error and other types of survey error. SDL error from synthesis tends to be smaller for numerical and financial variables than for categorical and demographic variables, while the opposite is true for the other sources of survey error. Additionally, SDL error tends to reduce estimated gaps in outcomes between sub-groups, whereas measurement error tends to increase estimated gaps. Our results highlight the importance of recognizing the presence of survey error already in the data before applying SDL and quantifying all possible sources of error as a way of evaluating and communicating data quality.

Future work should continue to explore the relationship between privacy protection and total survey error. Our results only speak to the accuracy of a single survey, for a limited set of variables, and for a particular synthesis model. Future work should assess to what extent our results generalize by applying the framework to more surveys, more variables, more administrative data sources, and more synthesis models. Statistical agencies may also benefit from incorporating this evaluation framework into the production process. Importantly, the framework does not require the use of synthesis for SDL and could be applied to surveys protected by other means.

Future work should also attempt to address some limitations of the current paper. One direction is to apply this framework to more statistics, such as variance or mean squared error. Additionally, while we focus on descriptive statistics, many users of microdata are interested

in model-based statistics. Although variable means and population sizes are important statistics used by many researchers and policymakers, SDL methods such as synthesis may impact the accuracy of these statistics in different ways from how they impact more complex statistics such as multi-variate modeled relationships. Evaluating the impact of SDL relative to other sources of survey error on modeled relationships is of great interest, but we leave this for future work.

Another direction is to address the fact that this framework treats the administrative data as if it were the truth. We know this is not always the case, but we argue that it is a more useful assumption than treating the survey data without SDL as if it were the truth and using that as the only evaluation benchmark. Future work that finds a way to relax the assumption that the administrative data are the truth would be valuable.

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. 2018. "The US Census Bureau adopts differential privacy." *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, 2867–2867.

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.

Abraham, Katharine G. 2019. "Reconciling data access and privacy: Building a sustainable model for the future." *AEA Papers and Proceedings*, 109: 409–413.

Agarwal, Anish, and Rahul Singh. 2024. "Causal Inference with Corrupted Data: Measurement Error, Missing Values, Discretization, and Differential Privacy." *arXiv preprint arXiv: 2107.02780*.

Alexander, J Trent, Michael Davern, and Betsey Stevenson. 2010. "The polls–Review: Inaccurate age and sex data in the census PUMS Files: Evidence and implications." *Public opinion quarterly*, 74(3): 551–569.

Barrientos, Andrés F, Aaron R Williams, Joshua Snoke, and Claire McKay Bowen. 2024. "A Feasibility Study of Differentially Private Summary Statistics and Regression Analyses with Evaluations on Administrative and Survey Data." *Journal of the American Statistical Association*, 119(545): 52–65.

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.

Bollinger, Christopher R, Barry T Hirsch, Charles M Hokayem, and James P Ziliak. 2019. "Trouble in the tails? What we know about earnings nonresponse 30 years after Lillard, Smith, and Welch." *Journal of Political Economy*, 127(5): 2143–2185.

Bound, John, Charles Brown, and Nancy Mathiowetz. 2001. "Measurement error in survey data." In *Handbook of Econometrics Vol. 5.* , ed. James J. Heckman and Edward Leamer, Chapter 59, 3705–3843. North-Holland.

Bowen, Claire McKay, Victoria L Bryant, Leonard Burman, Surachai Khitatrakun, Robert McClelland, Livia Mucciolo, Madeline Pickens, and Aaron R Williams. 2022. "Synthetic individual income tax data: promises and challenges." *National Tax Journal*, 75(4): 767–790.

Breiman, Leo, Jerome Friedman, R.A. Olshen, and Charles J. Stone. 1984. *Classification and regression trees.* Routledge.

Carr, Michael, Emily Wiemers, and Robert A Moffitt. 2023. "Using Synthetic Data to Estimate Earnings Dynamics: Evidence from the SIPP GSF and SIPP SSB." *Available at SSRN 4496224.*

Drechsler, Jörg, and Anna-Carolina Haensch. 2023. "30 years of synthetic data." *arXiv preprint arXiv:2304.02107.*

Garfinkel, Simson L, John M Abowd, and Sarah Powazek. 2018. "Issues encountered deploying differential privacy." *Proceedings of the 2018 Workshop on Privacy in the Electronic Society*, 133–137.

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.

Hawes, Michael B. 2020. "Implementing differential privacy: Seven lessons from the 2020 United States Census." *Harvard Data Science Review*, 2(2).

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 S. 2019. "Evolving measurement for an evolving economy: thoughts on 21st century US economic statistics." *Journal of Economic Perspectives*, 33(1): 165–184.

Kennickell, Arthur, and Julia Lane. 2006. "Measuring the impact of data protection techniques on data utility: Evidence from the survey of consumer finances." *International conference on privacy in statistical databases*, 291–303.

Klee, Mark A, Rebecca L Chenevert, and Kelly R Wilkin. 2019. "Revisiting the shape of earnings nonresponse." *Economics letters*, 184: 108663.

Komarova, Tatiana, and Denis Nekipelov. 2022. "Identification and Formal Privacy Guarantees." *Available at SSRN 3635824*.

Little, Roderick JA. 1993. "Statistical analysis of masked data." *Journal of official Statistics*, 9(2): 407–426.

Manski, Charles F. 2015. "Communicating uncertainty in official economic statistics: An appraisal fifty years after Morgenstern." *Journal of Economic Literature*, 53(3): 631–653.

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, Angela Wyse, and Kevin Corinth. 2023. "The size and census coverage of the US homeless population." *Journal of Urban Economics*, 136: 103559.

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.

Morgenstern, Oskar. 1963. *On the Accuracy of Economic Observations.* Princeton and London: Princeton University Press.

National Academies of Sciences, Engineering, and Medicine. 2023. *Assessing the 2020 Census: Final Report.* Washington, DC: The National Academies Press.

Rothbaum, Jonathan, and Adam Bee. 2021. "Coronavirus infects surveys, too: survey nonresponse bias and the Coronavirus pandemic." *US Census Bureau.*

Rubin, Donald B. 1993. "Statistical disclosure limitation." *Journal of official Statistics*, 9(2): 461–468.

Stanley, Jordan, and Evan Totty. 2024. "A Penny Synthesized is a Penny Earned? An Exploratory Analysis of Accuracy in the SIPP Synthetic Beta." *Harvard Data Science Review*, forthcoming.

U.S. Census Bureau. 2014. "American Community Survey: Design and Methodology."

U.S. Census Bureau. 2019. "ACS Accuracy of the Data (2019)." *https://www2.census.gov/programs-surveys/acs/tech$_d$ocs/pums/accuracy/*2019*AccuracyPUMS.pdf*.

U.S. Census Bureau. 2022. "Comparative Housing Characteristics." *https://data.census.gov/table/ACSCP1Y2022.CP04?q=Occupancy and Vacancy Status*, Accessed on 1 April 2024.

Vilhuber, Lars. 2020. "Reproducibility and replicability in economics." *Harvard Data Science Review*, 2(4).

Wagner, Deborah, and Mary Layne. 2014. "The Person Identification Validaiton System (PVS): Applying the Center for Administrative Records Research and Applicaitons' (CARRA) Record Linkage Software." *US Census Bureau.*

Appendix

A Data Sources, Cleaning, Linkage, and Synthesis

As discussed in section 2.1, some adjustments to the administrative data are required in order to align the administrative data with the survey target. First, any individuals/households excluded from the survey target must also be excluded from the administrative data. This exclusion criterion is less of an issue for the ACS than other Census Bureau surveys because the ACS samples from the entire residential population of the U.S. whereas other surveys (such as the Current Population Survey and Survey of Income and Program Participation) exclude group quarters (e.g., university student housing, worker's group living quarters, military barracks, retirement homes, correctional facilities, etc.). Consequently, the only individuals who should be excluded from the survey target are individuals who do not live in residential structures. This exclusion does not affect our analysis of residential structure variables (home value and property tax), but it does affect our analysis of individual-level variables. Currently we are unable to identify and exclude individuals not living in residential structures from the administrative data sources. There are ways in which we may be able to identify such individuals in future work, although evidence suggests the size of this group is quite small and thus unlikely to impact our current results (Meyer and Mittag, 2021*b*; Meyer, Wyse and Corinth, 2023).

There are some additional exclusions specific to the variables we analyzed. Wage and salary income and retirement income questions are limited to individuals age 15 or older in the survey. We do not observe age in the IRS data, so we have no direct way of excluding individuals with income below age 15 from the administrative data. We could link the IRS data to other administrative files in an attempt to identify such individuals, but this is likely a small group that would have little impact on our results. For residential structure variables, home value and property tax questions are limited to households that are owner-occupied. The Black Knight, Inc. data has information on owner-occupancy status, so we exclude

households from the Black Knight, Inc. data that are identified as not owner-occupied. The home value survey question is also in scope for households that are not owner-occupied but are vacant and either for sale or recently sold. The Black Knight, Inc. data does not provide vacancy information, so we are unable to directly identify equivalent houses in that data in order to ensure that such houses are not excluded from the survey target. However, non-rental vacancies make up less than 1% of households according to the ACS, so this is once again likely a small group that has little impact on our results (U.S. Census Bureau, 2022). Lastly, unlinkable records in the administrative data must be excluded.

[To be expanded upon in future drafts]

Table A1: Sources for the Variables of Interest

Variable	Administrative Data Source	Survey Question
Wage & Salary Income	IRS W-2 Forms	Person Question 43a: "Wages, salary, commissions, bonuses, or tips from all jobs. Report amount before deductions for taxes, bonds, dues, or other items." (Total amount for past 12 months)
Retirement Income	IRS 1099-R Forms	Person Question 43g: "Retirement income, pensions, survivor or disability income." (Include income from a previous employer or union, or any regular withdrawals or distributions from IRA, Roth IRA, 401(k), 403(b), or other accounts specifically designed for retirement. Do not include Social Security.)
Home Value	Black Knight, Inc.	Housing Question 19: "About how much do you think this house and lot, apartment, or mobile home (and lot, if owned) would sell for if it were for sale?" (Amount in dollars)
Property Tax	Black Knight, Inc.	Housing Question 20: "What are the annual real estate taxes on this property?" (Amount in dollars)
Birth Year	Social Security Administration	Person Question 4: "What is [your] age and what is [your] date of birth?" (Month, Day, Year of Birth)
Citizenship Status	Social Security Administration	Person Question 8: "Are [you] a citizen of the United States?" (Yes or No)
Race	U.S. Census Bureau	Person Question 6: "What is [your] race?" (Mark one or more boxes: White, Black or African Am., American Indian or Alaska Native, Asian Indian, Chinese, Filipino, Japanese, Korean, Vietnamese, Native Hawaiian, Guamanian or Chamorro, Somoan, Other Pacific Islander, Some other race)
Hispanic Status	U.S. Census Bureau	Person Question 5: "Are [you] of Hispanic, Latino, or Spanish origin?" (Yes or No)

Source: 2019 American Community Survey.

Table A2: Synthesis Details

Variable	Variable Type	Variable Values	Dependencies
Wage & Salary Income	Numerical	0–999,999	Age, Race, Hispanic Status, Female, Citizen, Education Level, Marital Status, Usual Weeks Worked, Usual Hours Worked Per Week, County
Retirement Income	Numerical	0–999,999	Age, Race, Hispanic Status, Female, Citizen, Education Level, Marital Status, Usual Weeks Worked, Usual Hours Worked Per Week, County
Home Value	Numerical	1,000–9,999,999	Householder Race, Householder Hispanic Status, Householder Age, Householder Sex, Householder Education Level, Householder Disability Status, Householder Poverty Status, Household Income, Acreage, Building Type, Housing Weight, Presence of a Mortgage, Number of Rooms, Number of Bedrooms, Number of Bathrooms, Kitchen, Running Water, Refrigerator, Sink, Stove, Pluming, Tenure, Vacancy Status, Year Built, County
Property Tax	Numerical	0–99,999	*Synthetic Home Value*, Military Disability Status, County
Birth Year	Numerical		Female, School Enrollment, School Grade, Education Level, Marital Status, Race, Hispanic Status, Relationship to Householder, Usual Weeks Worked, Usual Hours Worked Per Week, Wage and Salary Income, Retirement Income, Social Security Income, Public Assistance Income, Self-Employment Income, Total Income, County
Citizenship Status	Categorical	1–5	Age, Female, Education Level, Marital Status, Race, Hispanic Status, Usual Weeks Worked, Usual Hours Worked Per Week, Wage and Salary Income, Retirement Income, Social Security Income, Public Assistance Income, Self-Employment Income, Total Income, County
Race	Categorical	1–7	Age, Female, Citizenship Status, Education Level, Marital Status, Usual Weeks Worked, Usual Hours Worked Per Week, Wage and Salary Income, Retirement Income, Social Security Income, Public Assistance Income, Self-Employment Income, Total Income, County
Hispanic Status	Categorical	1–2	*Synthetic Race*, Age, Female, Citizenship Status, Education Level, Marital Status, Usual Weeks Worked, Usual Hours Worked Per Week, Wage and Salary Income, Retirement Income, Social Security Income, Public Assistance Income, Self-Employment Income, Total Income, County

Source: 2019 American Community Survey.
Note: Additional synthesis details: (1) each variable was synthesized using state as an independent feature, meaning that a different tree was trained for each state; (2) the only privacy parameter used for the trees was a minimum leaf size of 5; synthesis was performed five separate times in order to create five implicates of synthetic data for each variable.

B Error Misclassification

Table B1: Total Survey Error Misclassification Breakdown for Numeric Variables

	(1) Wage & Salary	(2) Retirement Inc.	(3) Home Value	(4) Prop. Taxes	(5) Birth Year
Measurement Error	$4,237	-$29.53	$50,460	$617.10	-0.039
Correctly Classified	$1,034	$283.80	-$311.20	$22.92	-0.039
False Negative	-$1,036	-$871.70	–	-$20.91	–
False Positive	$4,239	$558.40	$50,770	$615.10	–
Non-response Error	$1,450	$136.20	$11,110	$116.30	0.01
Correctly Classified	$337.40	$57.05	$1,526	-$4.02	0.01
False Negative	-$431.30	-$186.50	–	-$8.76	–
False Positive	$1,544	$265.60	$9,581	$129.10	–
SDL Error	$207.10	$149.90	$1,060	-$5.23	-0.992
Correctly Classified	$203.80	$11.89	$1,060	-$8.82	-0.992
False Negative	-$1,373	$3,349	–	-$82.84	–
False Positive	$1,377	$1,589	–	$86.42	–

Source: 2019 American Community Survey (ACS), Census Bureau's Best Race and Ethnicity internal file, Social Security Administration records.

Note: The Statistical Disclosure Limitation (SDL) method being applied here is synthesis applied to the ACS variable listed in the column heading using classification and regression tree methods to replace observed survey values with modeled values. Correctly Classified corresponds to observations with non-missing positive values in both the survey and administrative data and reports the weighted average difference (survey value minus administrative value). False Negative corresponds to non-missing values of $0 in the survey and non-missing positive values in the administrative data and reports the weighted average difference (survey minus administrative). False Positive corresponds to observations with non-missing positive values in the survey and non-missing or implied values of $0 in the administrative data and reports the weighted average difference (survey minus administrative). See equations 4a-4c in section 2.2 for more details. Home value is missing false negative for measurement error and non-response error (and missing both false negative and false positive for SDL error) because home value is bottom-coded at $1,000 in the survey but not bottom-coded in the administrative data. Birth year is missing false negative and false positive because birth year cannot have a zero value. All results have been reviewed by the U.S. Census Bureau to ensure no confidential information has been disclosed: CBDRB-FY24-CED010-0001, CBDRB-FY24-CED010-0002.

Table B2: Total Survey Error Misclassification Breakdown for Categorical Variables

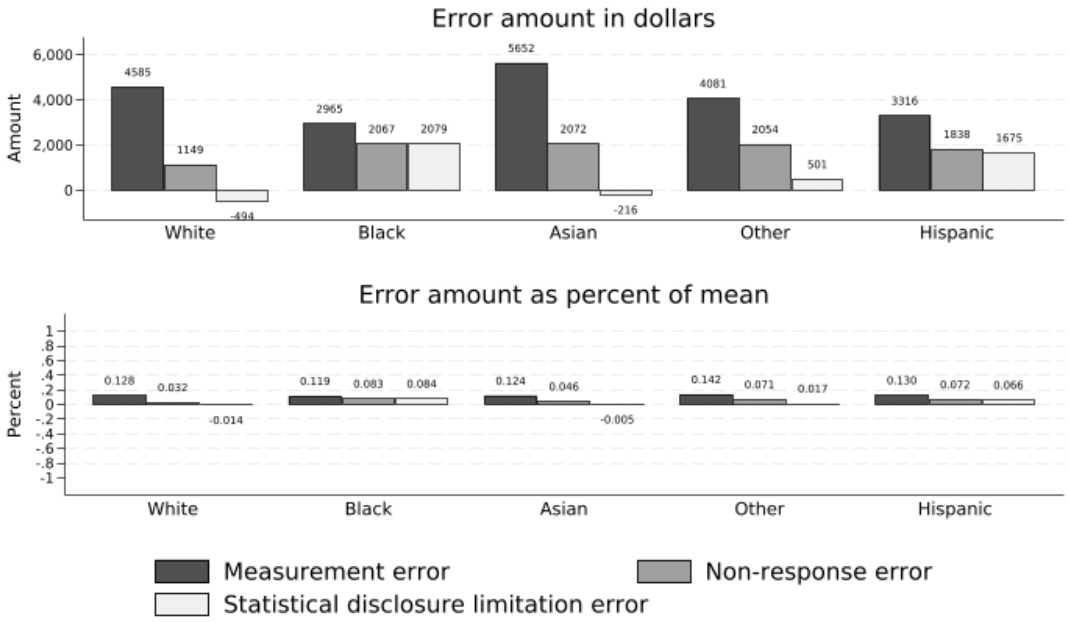
	(1) White	(2) Black	(3) Asian	(4) Other Race	(5) Hispanic Status	(6) Citizenship
Measurement Error	0.0090	0.0010	0.0010	-0.0030	-0.0080	0.0250
Correctly Classified	–	–	–	–	–	–
False Negative	-0.0012	-0.0016	-0.0003	-0.0044	-0.0090	-0.0061
False Positive	0.0103	0.0023	0.0011	0.0016	0.0013	0.0308
Non-response Error	0.0010	-0.0005	-0.0001	-0.0002	0.0000	0.0030
Correctly Classified	–	–	–	–	–	–
False Negative	-0.0006	-0.0007	-0.0003	-0.0004	-0.0005	-0.0028
False Positive	0.0014	0.0002	0.0002	0.0002	0.0006	0.0059
SDL Error	0.0270	-0.0180	0.0020	-0.0010	-0.0100	0.0020
Correctly Classified	–	–	–	–	–	–
False Negative	-0.0963	-0.0647	-0.0278	-0.0259	-0.0748	-0.0394
False Positive	0.1234	0.0467	0.0299	0.0254	0.0641	0.0411

Source: 2019 American Community Survey (ACS), Census Bureau's Best Race and Ethnicity internal file, Social Security Administration records.

Note: The Statistical Disclosure Limitation (SDL) method being applied here is synthesis applied to the ACS variable listed in the column heading using classification and regression tree methods to replace observed survey values with modeled values. Correctly Classified corresponds to observations with the same categorical value in the survey and administrative data. False Negative corresponds to observations with a non-missing value of 0 for the given category indicator in the survey and a value of 1 in the administrative data and reports the weighted average difference (survey minus administrative). False Positive corresponds to observations with a non-missing value of 1 for the given category indicator in the survey and a value of 0 in the administrative data and reports the weighted average difference (survey minus administrative). See equations 4a-4c in section 2.2 for more details. Note that correctly classified is always missing because the average difference between survey and administrative records for correctly classified binary categories is zero. All results have been reviewed by the U.S. Census Bureau to ensure no confidential information has been disclosed: CBDRB-FY24-CED010-0001.

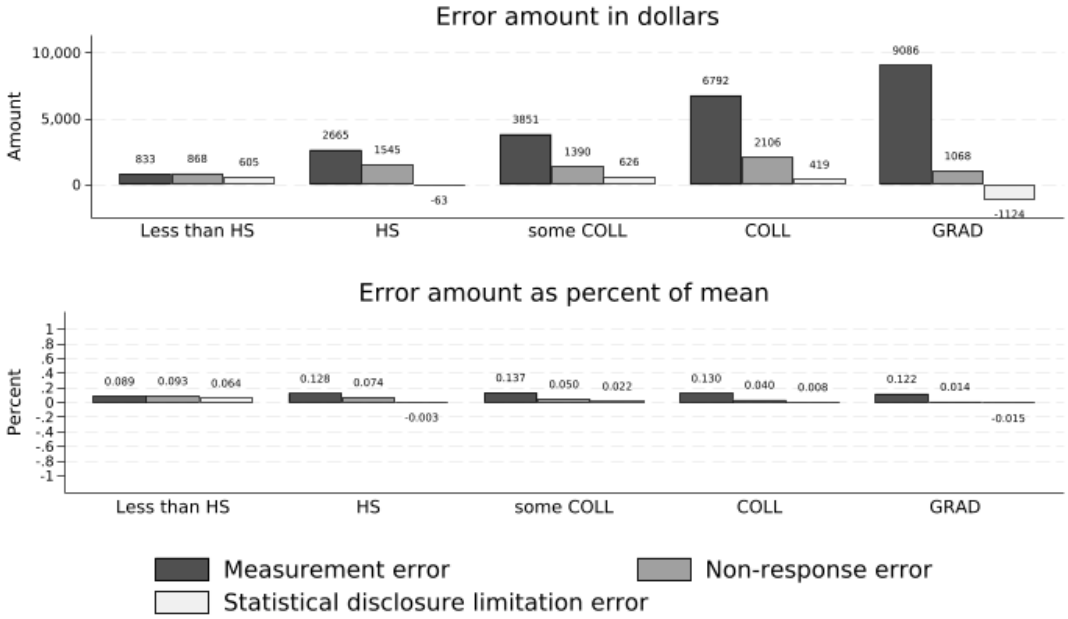
C Sub-Group Errors

Figure C1: Wage and Salary Income by Race



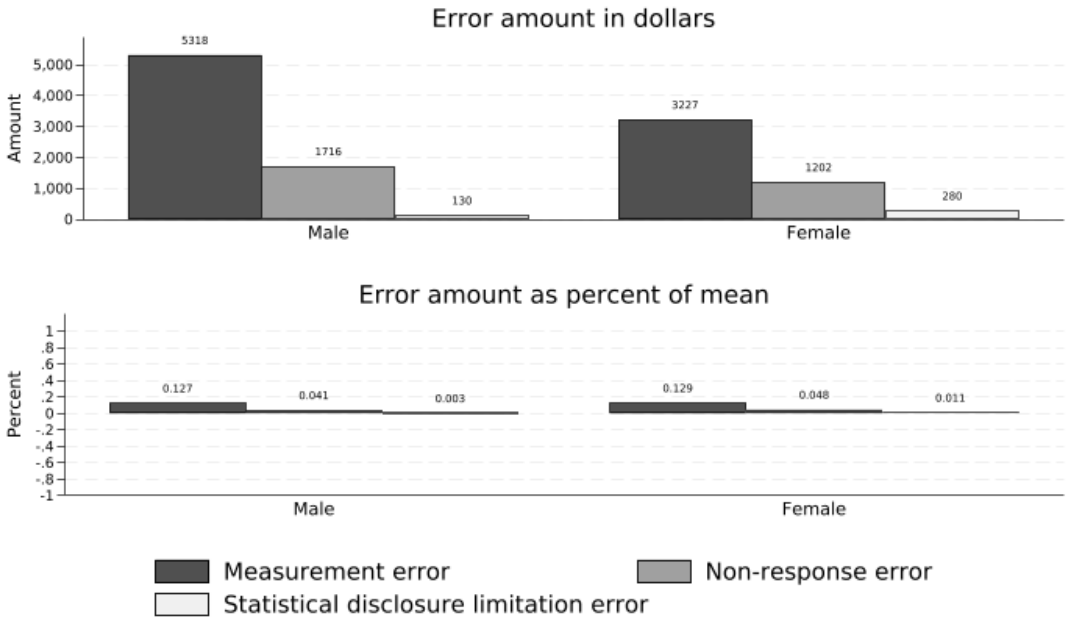
Source: 2019 American Community Survey and IRS W-2 forms for tax year 2018.
Note: The top figure reports error amounts by error type and by race. The bottom figure expresses the amounts from the top figure as a percentage of the survey-based mean wage and salary income by race. See Section 2.1 for more information on how the error amounts are computed and Section 4.2 for more information on how error amounts are expressed as a percentage of group-specific means. All results have been reviewed by the U.S. Census Bureau to ensure no confidential information has been disclosed: CBDRB-FY24-CED010-0001.

Figure C2: Wage and Salary Income by Education



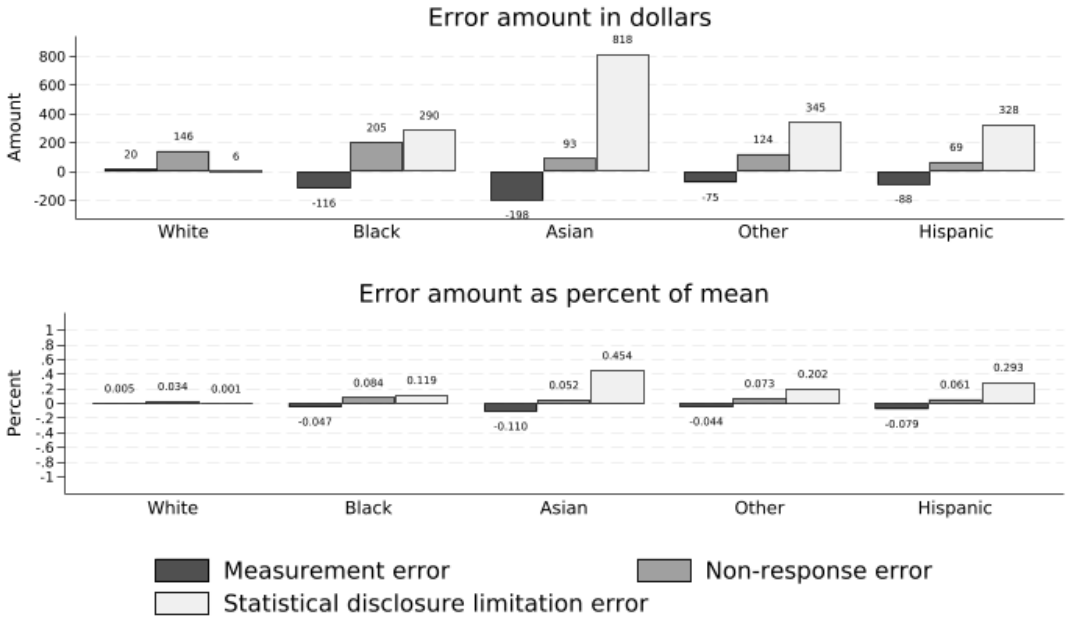
Source: 2019 American Community Survey and IRS W-2 forms for tax year 2018.
Note: The top figure reports error amounts by error type and by education. The bottom figure expresses the amounts from the top figure as a percentage of the survey-based mean wage and salary income by education. See Section 2.1 for more information on how the error amounts are computed and Section 4.2 for more information on how error amounts are expressed as a percentage of group-specific means. All results have been reviewed by the U.S. Census Bureau to ensure no confidential information has been disclosed: CBDRB-FY24-CED010-0001.

Figure C3: Wage and Salary Income by Sex



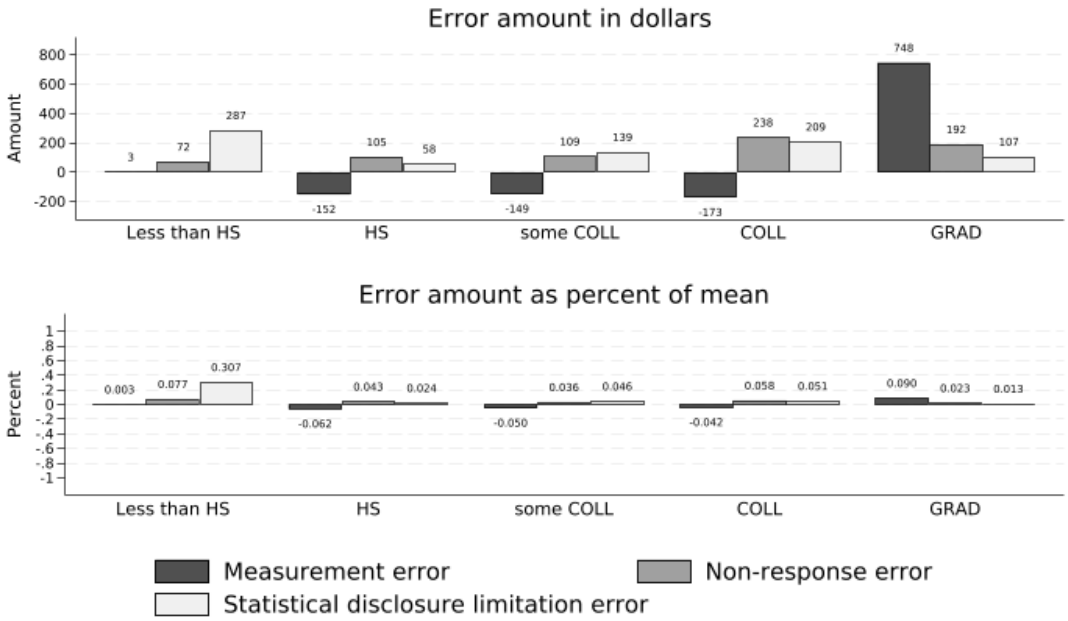
Source: 2019 American Community Survey and IRS W-2 forms for tax year 2018.
Note: The top figure reports error amounts by error type and by sex. The bottom figure expresses the amounts from the top figure as a percentage of the survey-based mean wage and salary income by sex. See Section 2.1 for more information on how the error amounts are computed and Section 4.2 for more information on how error amounts are expressed as a percentage of group-specific means. All results have been reviewed by the U.S. Census Bureau to ensure no confidential information has been disclosed: CBDRB-FY24-CED010-0001.

Figure C4: Retirement Income by Race



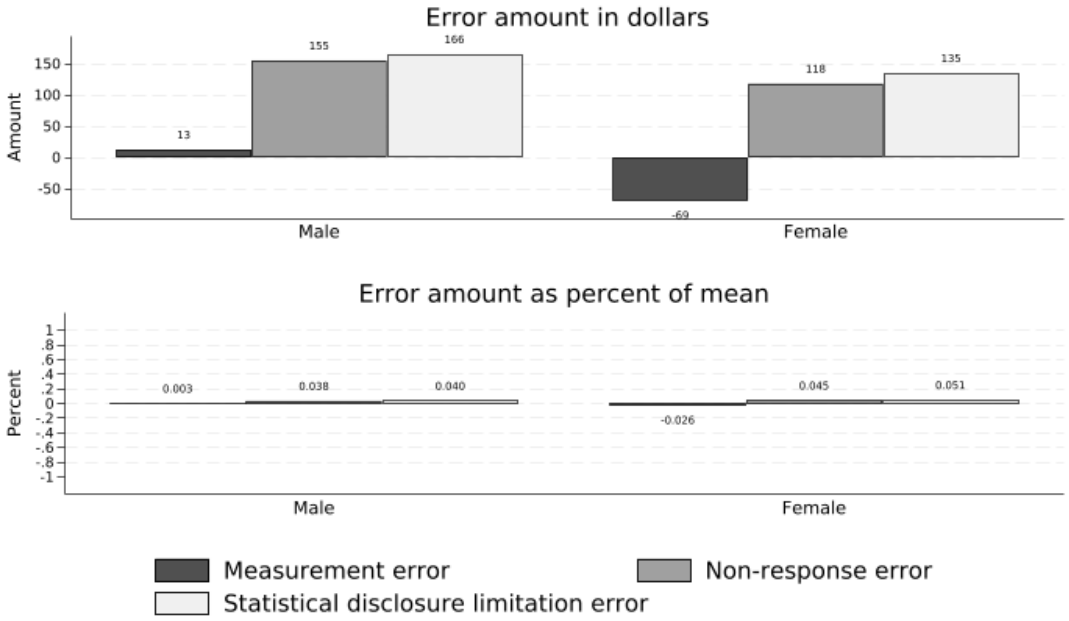
Source: 2019 American Community Survey and IRS 1099-R forms for tax year 2018.
Note: The top figure reports error amounts by error type and by race. The bottom figure expresses the amounts from the top figure as a percentage of the survey-based mean retirement income by race. See Section 2.1 for more information on how the error amounts are computed and Section 4.2 for more information on how error amounts are expressed as a percentage of group-specific means. All results have been reviewed by the U.S. Census Bureau to ensure no confidential information has been disclosed: CBDRB-FY24-CED010-0001.

Figure C5: Retirement Income by Education



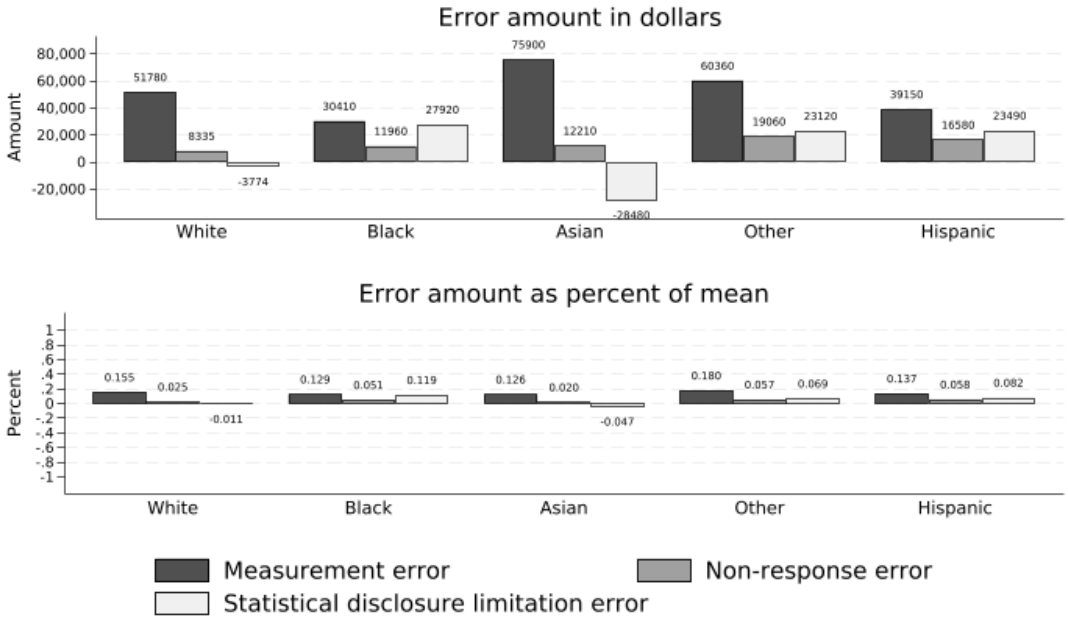
Source: 2019 American Community Survey and IRS 1099-R forms for tax year 2018.
Note: The top figure reports error amounts by error type and by education. The bottom figure expresses the amounts from the top figure as a percentage of the survey-based mean retirement income by education. See Section 2.1 for more information on how the error amounts are computed and Section 4.2 for more information on how error amounts are expressed as a percentage of group-specific means. All results have been reviewed by the U.S. Census Bureau to ensure no confidential information has been disclosed: CBDRB-FY24-CED010-0001.

Figure C6: Retirement Income by Sex



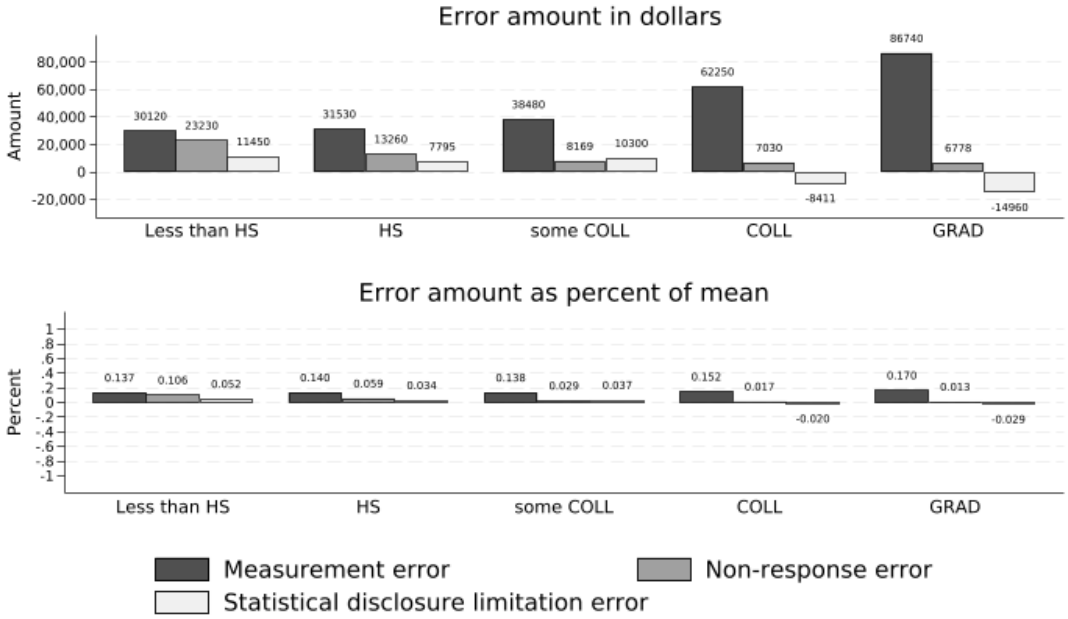
Source: 2019 American Community Survey and IRS 1099-R forms for tax year 2018.
Note: The top figure reports error amounts by error type and by sex. The bottom figure expresses the amounts from the top figure as a percentage of the survey-based mean retirement income by sex. See Section 2.1 for more information on how the error amounts are computed and Section 4.2 for more information on how error amounts are expressed as a percentage of group-specific means. All results have been reviewed by the U.S. Census Bureau to ensure no confidential information has been disclosed: CBDRB-FY24-CED010-0001.

Figure C7: Home Value by Race



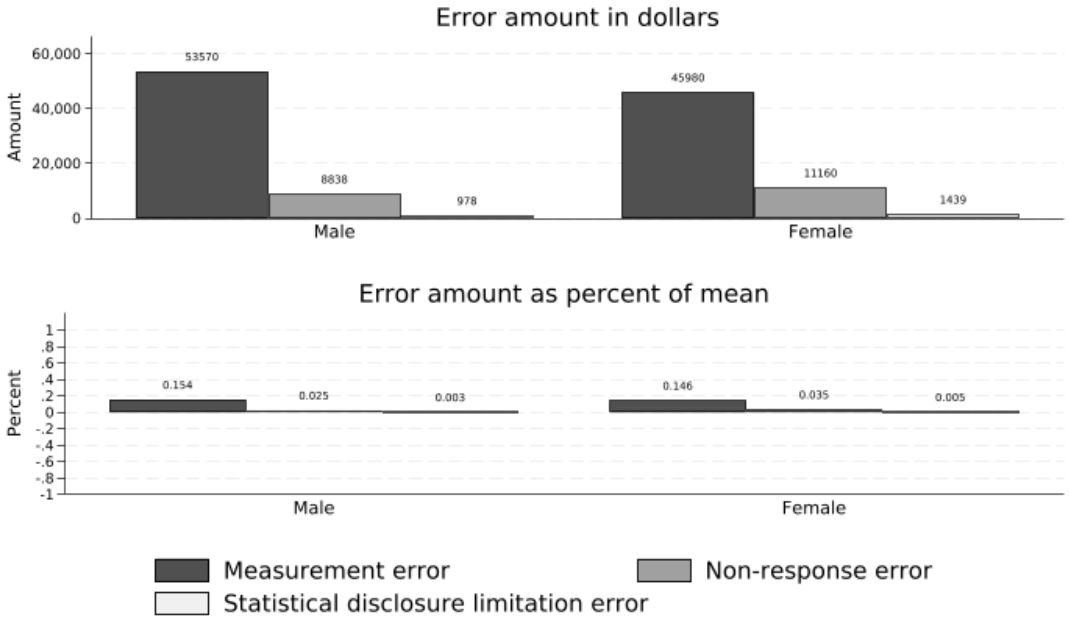
Source: 2019 American Community Survey and Black Knight, Inc. home valuation data.
Note: The top figure reports error amounts by error type and by race. The bottom figure expresses the amounts from the top figure as a percentage of the survey-based mean home value by race. See Section 2.1 for more information on how the error amounts are computed and Section 4.2 for more information on how error amounts are expressed as a percentage of group-specific means. All results have been reviewed by the U.S. Census Bureau to ensure no confidential information has been disclosed: CBDRB-FY24-CED010-0001.

Figure C8: Home Value by Education



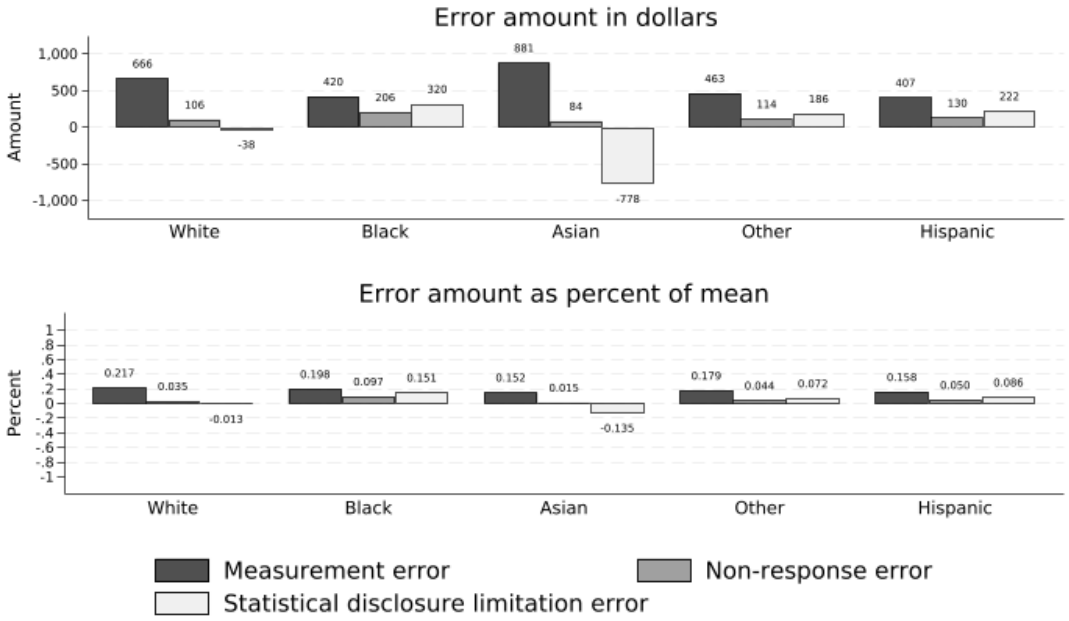
Source: 2019 American Community Survey and Black Knight, Inc. home valuation data.
Note: The top figure reports error amounts by error type and by education. The bottom figure expresses the amounts from the top figure as a percentage of the survey-based mean home value by education. See Section 2.1 for more information on how the error amounts are computed and Section 4.2 for more information on how error amounts are expressed as a percentage of group-specific means. All results have been reviewed by the U.S. Census Bureau to ensure no confidential information has been disclosed: CBDRB-FY24-CED010-0001.

Figure C9: Home Value by Sex



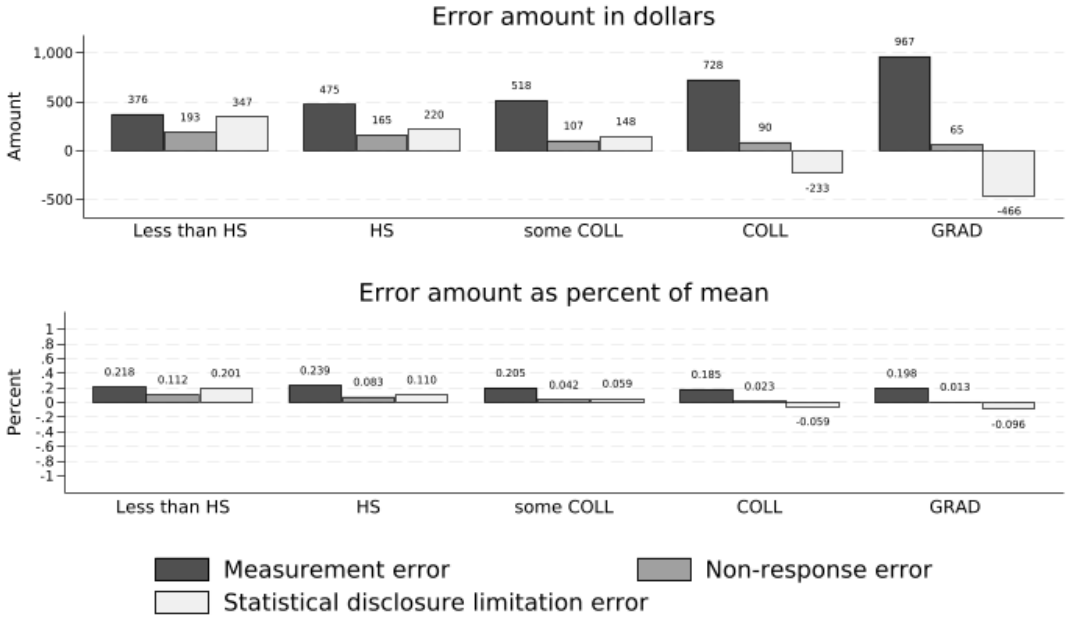
Source: 2019 American Community Survey and Black Knight, Inc. home valuation data.
Note: The top figure reports error amounts by error type and by sex. The bottom figure expresses the amounts from the top figure as a percentage of the survey-based mean home value by sex. See Section 2.1 for more information on how the error amounts are computed and Section 4.2 for more information on how error amounts are expressed as a percentage of group-specific means. All results have been reviewed by the U.S. Census Bureau to ensure no confidential information has been disclosed: CBDRB-FY24-CED010-0001.

Figure C10: Property Taxes by Race



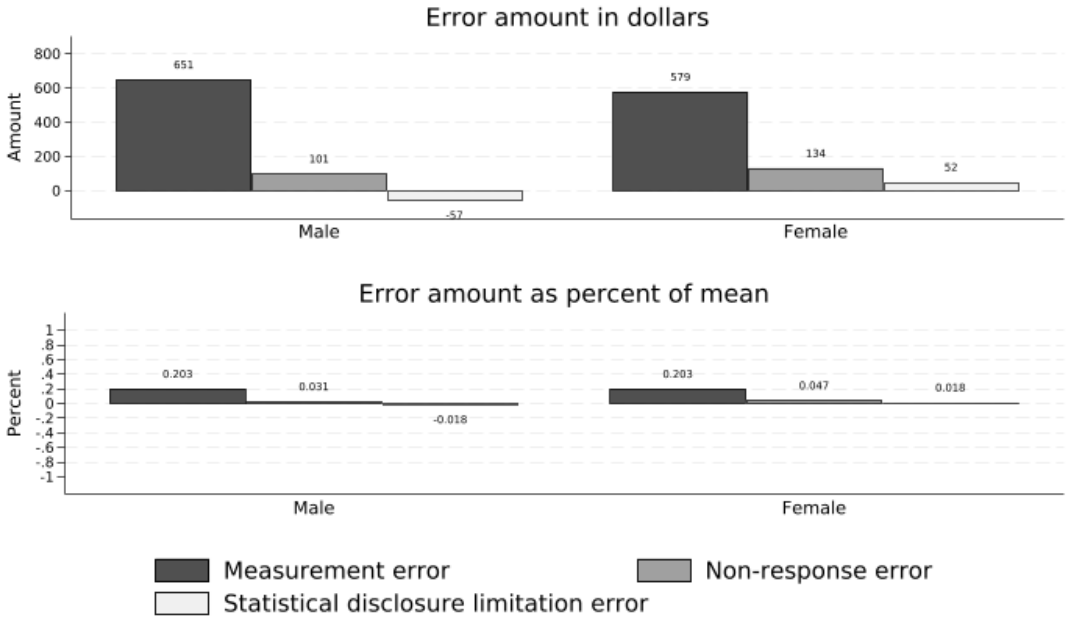
Source: 2019 American Community Survey and Black Knight, Inc. real estate records.
Note: The top figure reports error amounts by error type and by race. The bottom figure expresses the amounts from the top figure as a percentage of the survey-based mean property taxes by race. See Section 2.1 for more information on how the error amounts are computed and Section 4.2 for more information on how error amounts are expressed as a percentage of group-specific means. All results have been reviewed by the U.S. Census Bureau to ensure no confidential information has been disclosed: CBDRB-FY24-CED010-0002.

Figure C11: Property Taxes by Education



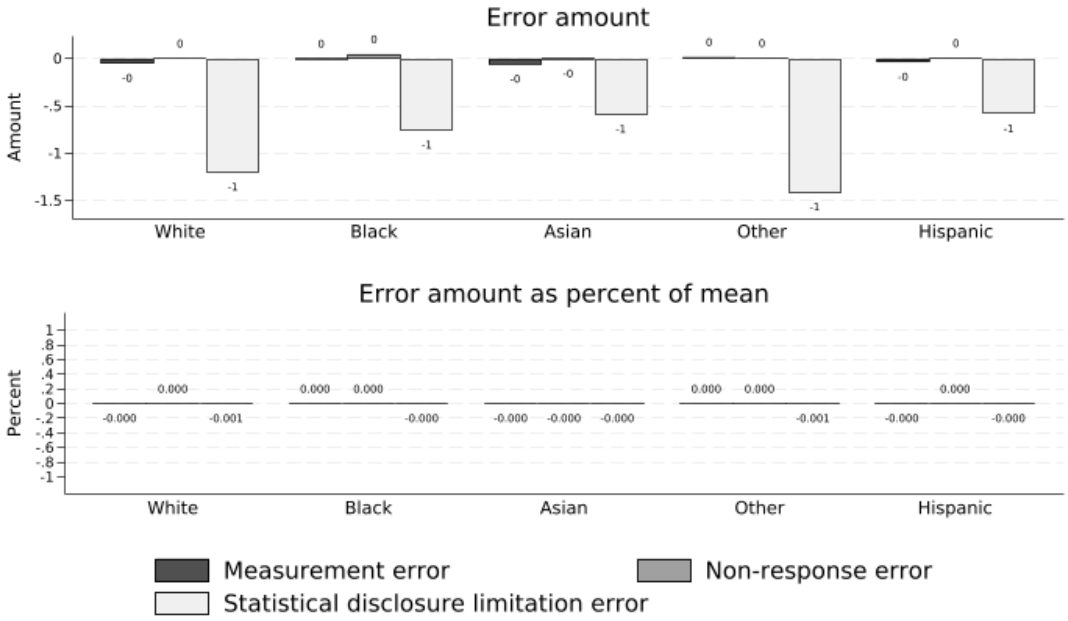
Source: 2019 American Community Survey and Black Knight, Inc. real estate data.
Note: The top figure reports error amounts by error type and by education. The bottom figure expresses the amounts from the top figure as a percentage of the survey-based mean property taxes by education. See Section 2.1 for more information on how the error amounts are computed and Section 4.2 for more information on how error amounts are expressed as a percentage of group-specific means. All results have been reviewed by the U.S. Census Bureau to ensure no confidential information has been disclosed: CBDRB-FY24-CED010-0002.

Figure C12: Property Taxes by Sex



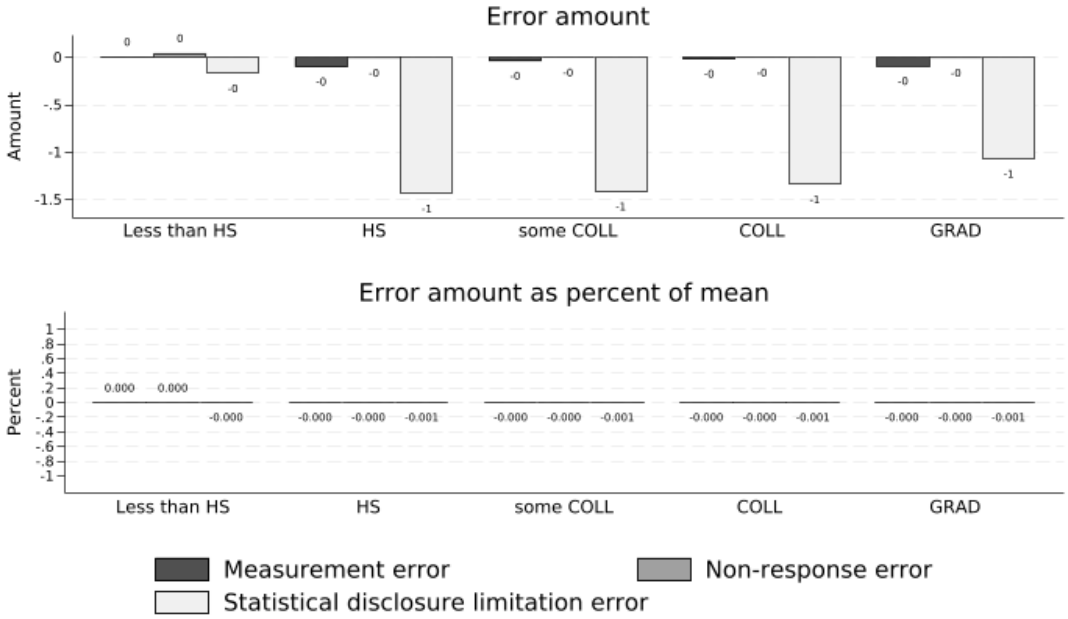
Source: 2019 American Community Survey and Black Knight, Inc. real estate records.
Note: The top figure reports error amounts by error type and by sex. The bottom figure expresses the amounts from the top figure as a percentage of the survey-based mean property taxes by sex. See Section 2.1 for more information on how the error amounts are computed and Section 4.2 for more information on how error amounts are expressed as a percentage of group-specific means. All results have been reviewed by the U.S. Census Bureau to ensure no confidential information has been disclosed: CBDRB-FY24-CED010-0002.

Figure C13: Birth Year by Race



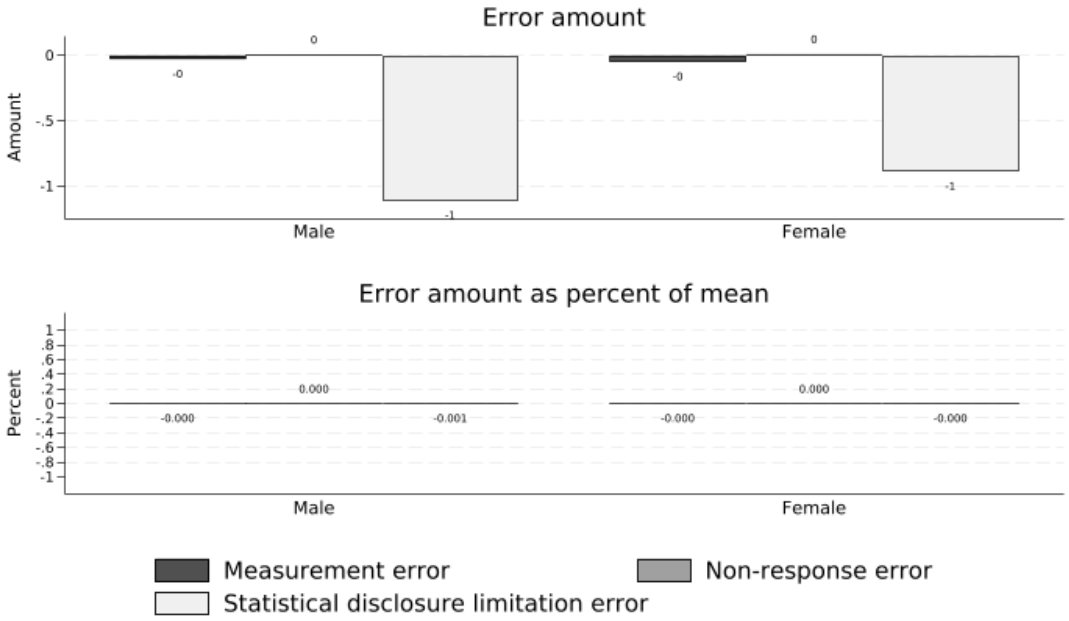
Source: 2019 American Community Survey and Social Security Administration records.
Note: The top figure reports error amounts by error type and by race. The bottom figure expresses the amounts from the top figure as a percentage of the survey-based mean birth year by race. See Section 2.1 for more information on how the error amounts are computed and Section 4.2 for more information on how error amounts are expressed as a percentage of group-specific means. All results have been reviewed by the U.S. Census Bureau to ensure no confidential information has been disclosed: CBDRB-FY24-CED010-0001.

Figure C14: Birth Year by Education



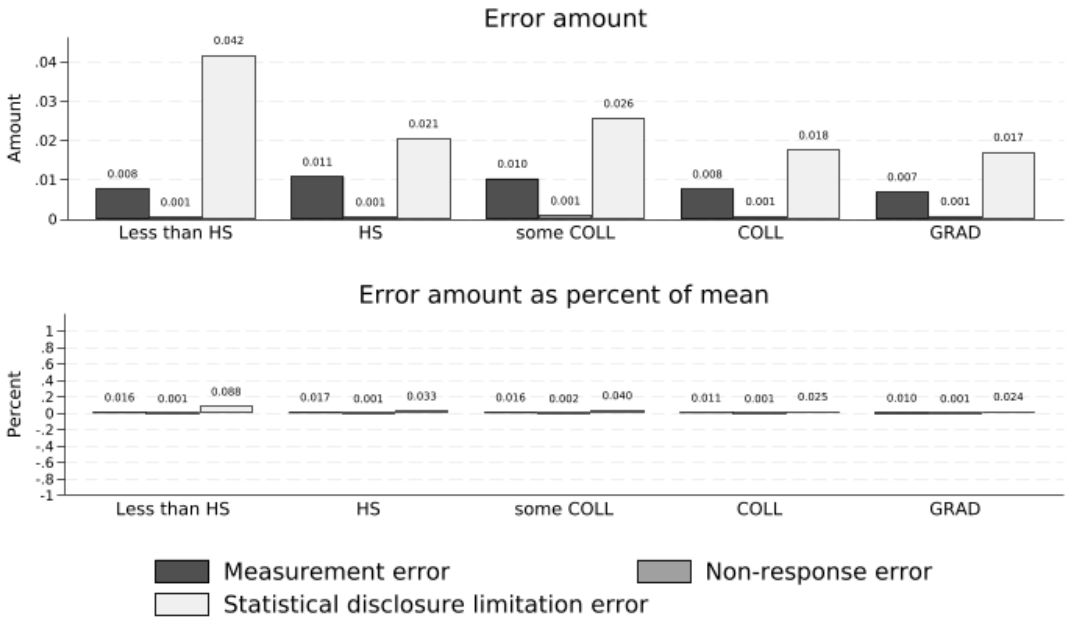
Source: 2019 American Community Survey and Social Security Administration records.
Note: The top figure reports error amounts by error type and by education. The bottom figure expresses the amounts from the top figure as a percentage of the survey-based mean birth year by education. See Section 2.1 for more information on how the error amounts are computed and Section 4.2 for more information on how error amounts are expressed as a percentage of group-specific means. All results have been reviewed by the U.S. Census Bureau to ensure no confidential information has been disclosed: CBDRB-FY24-CED010-0001.

Figure C15: Birth Year by Sex



Source: 2019 American Community Survey and Social Security Administration records.
Note: The top figure reports error amounts by error type and by sex. The bottom figure expresses the amounts from the top figure as a percentage of the survey-based mean birth year by sex. See Section 2.1 for more information on how the error amounts are computed and Section 4.2 for more information on how error amounts are expressed as a percentage of group-specific means. All results have been reviewed by the U.S. Census Bureau to ensure no confidential information has been disclosed: CBDRB-FY24-CED010-0001.

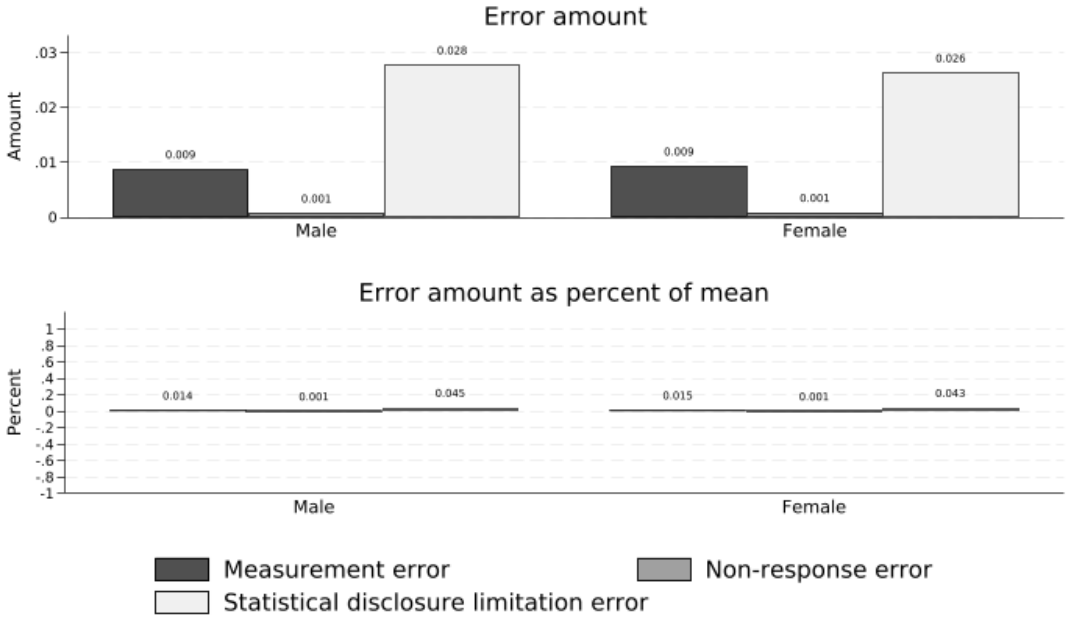
Figure C16: Proportion White by Education



Source: 2019 American Community Survey and Census Bureau's Best Race and Ethnicity internal file.

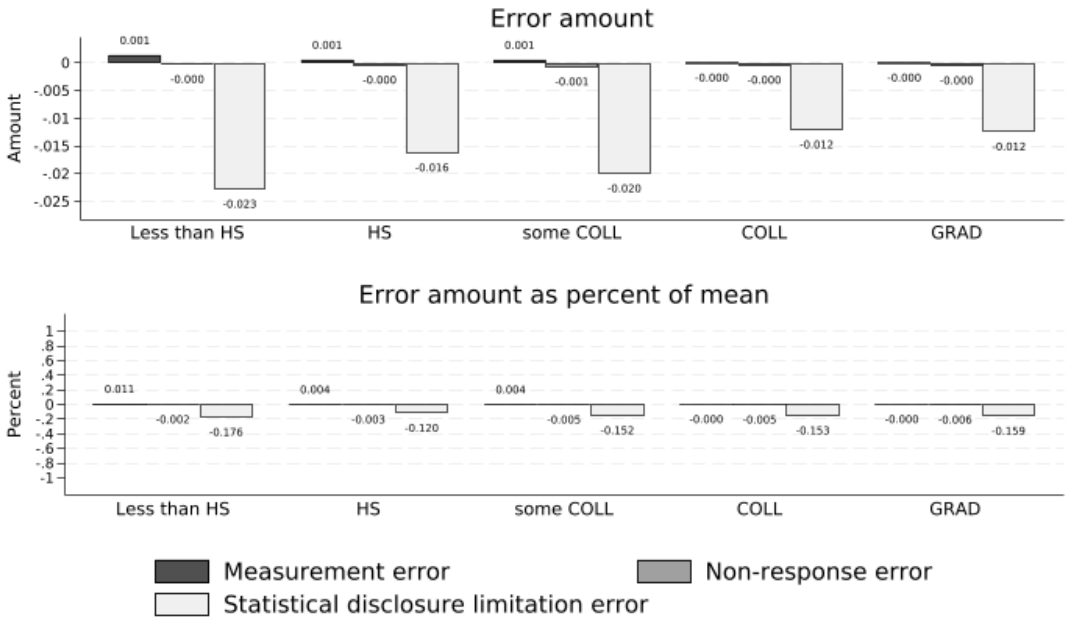
Note: The top figure reports error amounts by error type and by education. The bottom figure expresses the amounts from the top figure as a percentage of the survey-based mean proportion White by education. See Section 2.1 for more information on how the error amounts are computed and Section 4.2 for more information on how error amounts are expressed as a percentage of group-specific means. All results have been reviewed by the U.S. Census Bureau to ensure no confidential information has been disclosed: CBDRB-FY24-CED010-0001.

Figure C17: Proportion White by Sex



Source: 2019 American Community Survey and Census Bureau's Best Race and Ethnicity internal file.
Note: The top figure reports error amounts by error type and by sex. The bottom figure expresses the amounts from the top figure as a percentage of the survey-based mean proportion White by sex. See Section 2.1 for more information on how the error amounts are computed and Section 4.2 for more information on how error amounts are expressed as a percentage of group-specific means. All results have been reviewed by the U.S. Census Bureau to ensure no confidential information has been disclosed: CBDRB-FY24-CED010-0001.

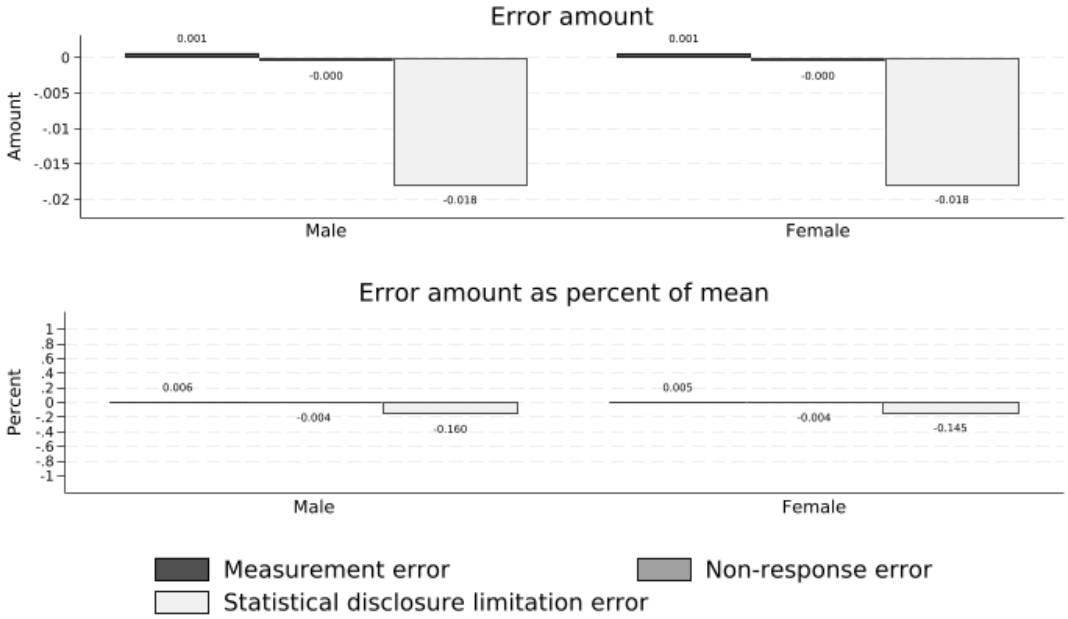
Figure C18: Proportion Black by Education



Source: 2019 American Community Survey and Census Bureau's Best Race and Ethnicity internal file.

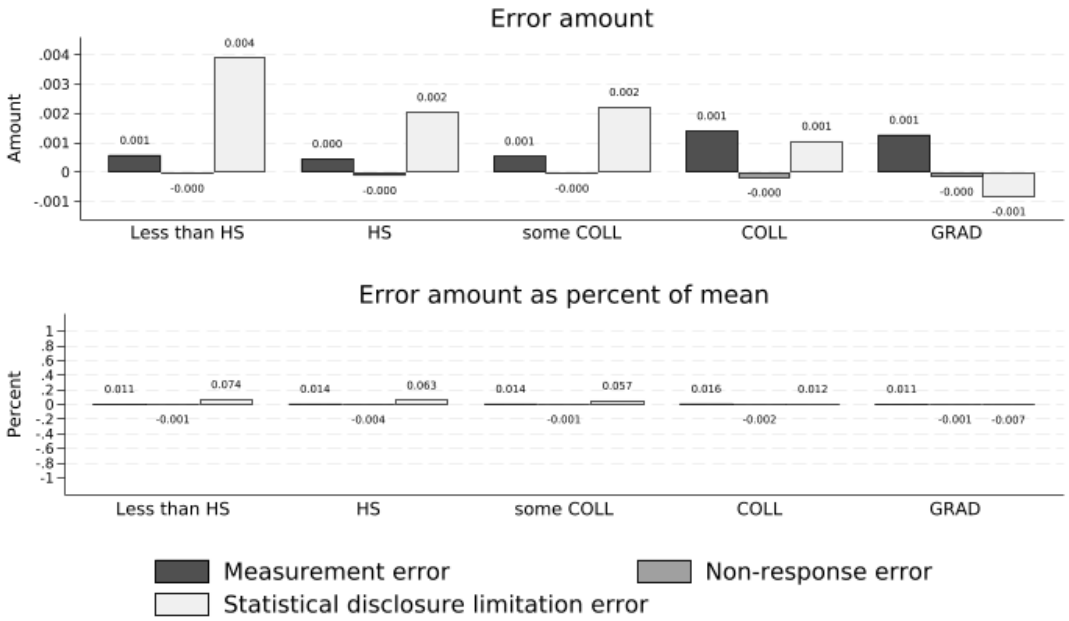
Note: The top figure reports error amounts by error type and by education. The bottom figure expresses the amounts from the top figure as a percentage of the survey-based mean proportion Black by education. See Section 2.1 for more information on how the error amounts are computed and Section 4.2 for more information on how error amounts are expressed as a percentage of group-specific means. All results have been reviewed by the U.S. Census Bureau to ensure no confidential information has been disclosed: CBDRB-FY24-CED010-0001.

Figure C19: Proportion Black by Sex



Source: 2019 American Community Survey and Census Bureau's Best Race and Ethnicity internal file.
Note: The top figure reports error amounts by error type and by sex. The bottom figure expresses the amounts from the top figure as a percentage of the survey-based mean proportion Black by sex. See Section 2.1 for more information on how the error amounts are computed and Section 4.2 for more information on how error amounts are expressed as a percentage of group-specific means. All results have been reviewed by the U.S. Census Bureau to ensure no confidential information has been disclosed: CBDRB-FY24-CED010-0001.

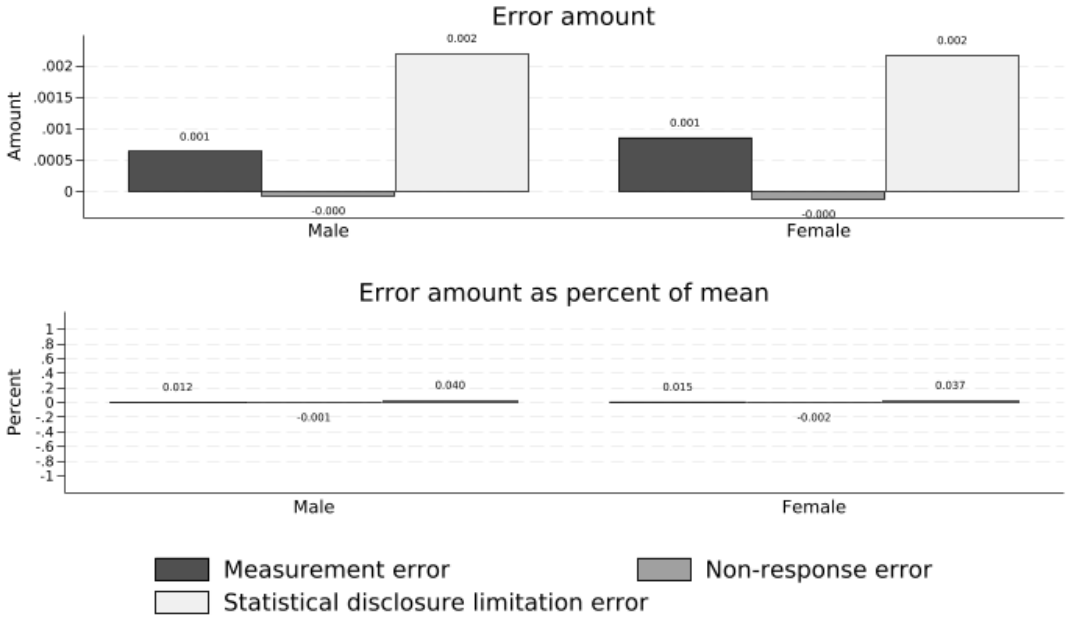
Figure C20: Proportion Asian by Education



Source: 2019 American Community Survey and Census Bureau's Best Race and Ethnicity internal file.

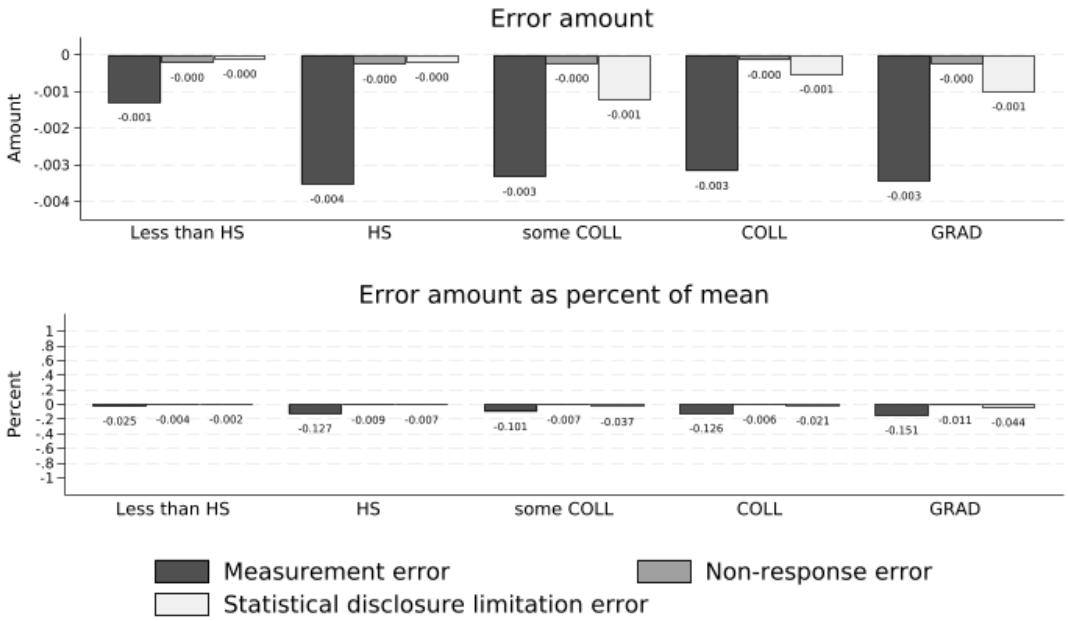
Note: The top figure reports error amounts by error type and by education. The bottom figure expresses the amounts from the top figure as a percentage of the survey-based mean proportion Asian by education. See Section 2.1 for more information on how the error amounts are computed and Section 4.2 for more information on how error amounts are expressed as a percentage of group-specific means. All results have been reviewed by the U.S. Census Bureau to ensure no confidential information has been disclosed: CBDRB-FY24-CED010-0001.

Figure C21: Proportion Asian by Sex



Source: 2019 American Community Survey and Census Bureau's Best Race and Ethnicity internal file.
Note: The top figure reports error amounts by error type and by sex. The bottom figure expresses the amounts from the top figure as a percentage of the survey-based mean proportion Asian by sex. See Section 2.1 for more information on how the error amounts are computed and Section 4.2 for more information on how error amounts are expressed as a percentage of group-specific means. All results have been reviewed by the U.S. Census Bureau to ensure no confidential information has been disclosed: CBDRB-FY24-CED010-0001.

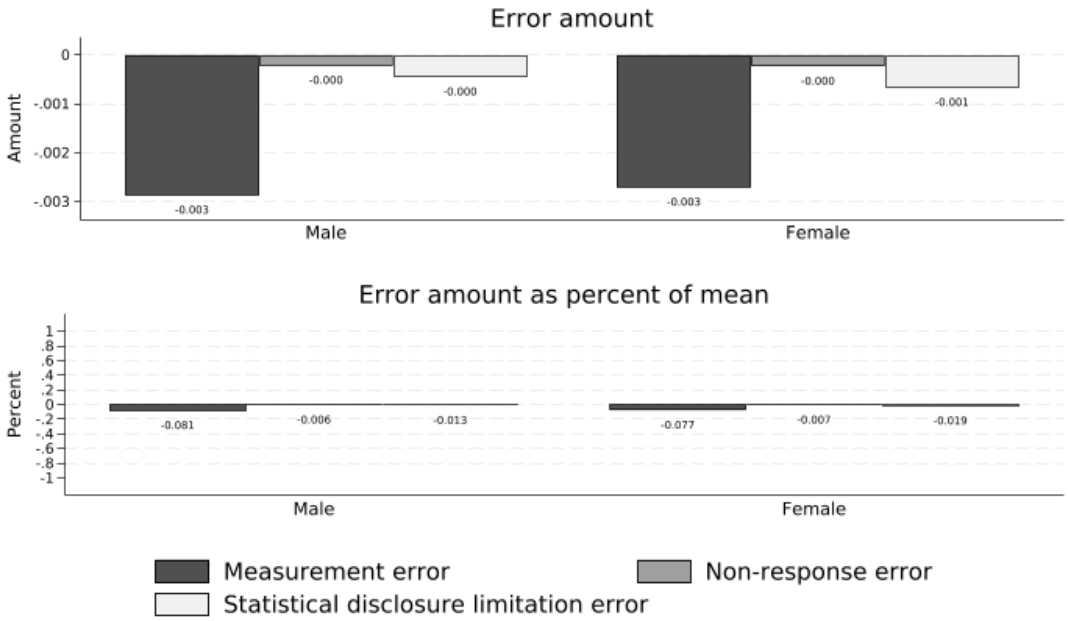
Figure C22: Proportion Other Race by Education



Source: 2019 American Community Survey and Census Bureau's Best Race and Ethnicity internal file.

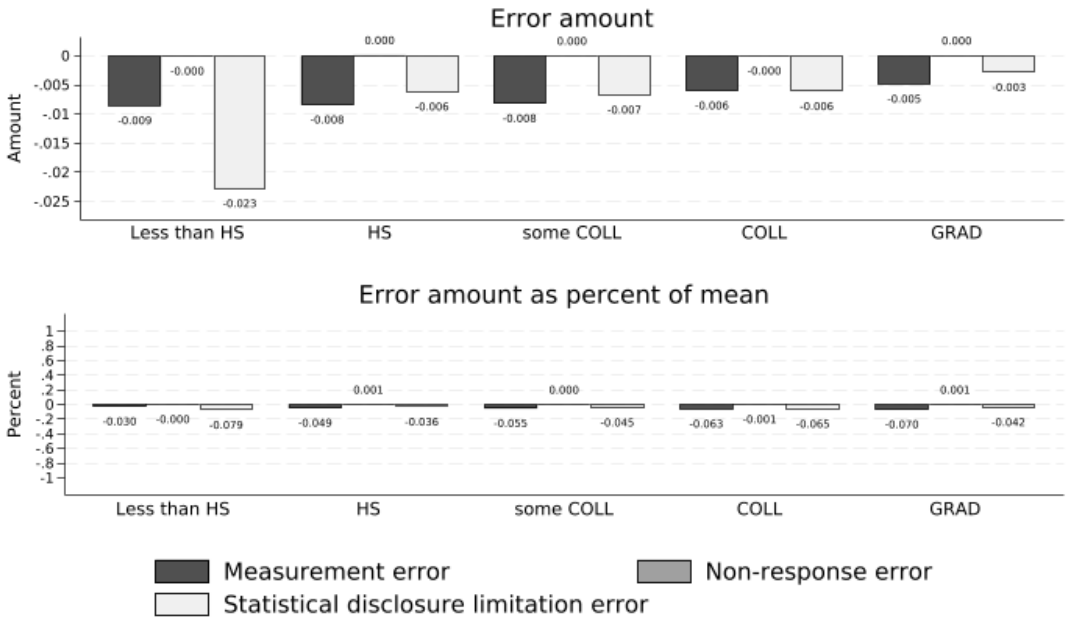
Note: The top figure reports error amounts by error type and by education. The bottom figure expresses the amounts from the top figure as a percentage of the survey-based mean proportion other race by education. See Section 2.1 for more information on how the error amounts are computed and Section 4.2 for more information on how error amounts are expressed as a percentage of group-specific means. All results have been reviewed by the U.S. Census Bureau to ensure no confidential information has been disclosed: CBDRB-FY24-CED010-0001.

Figure C23: Proportion Other Race by Sex



Source: 2019 American Community Survey and Census Bureau's Best Race and Ethnicity internal file.
Note: The top figure reports error amounts by error type and by sex. The bottom figure expresses the amounts from the top figure as a percentage of the survey-based mean proportion other race by sex. See Section 2.1 for more information on how the error amounts are computed and Section 4.2 for more information on how error amounts are expressed as a percentage of group-specific means. All results have been reviewed by the U.S. Census Bureau to ensure no confidential information has been disclosed: CBDRB-FY24-CED010-0001.

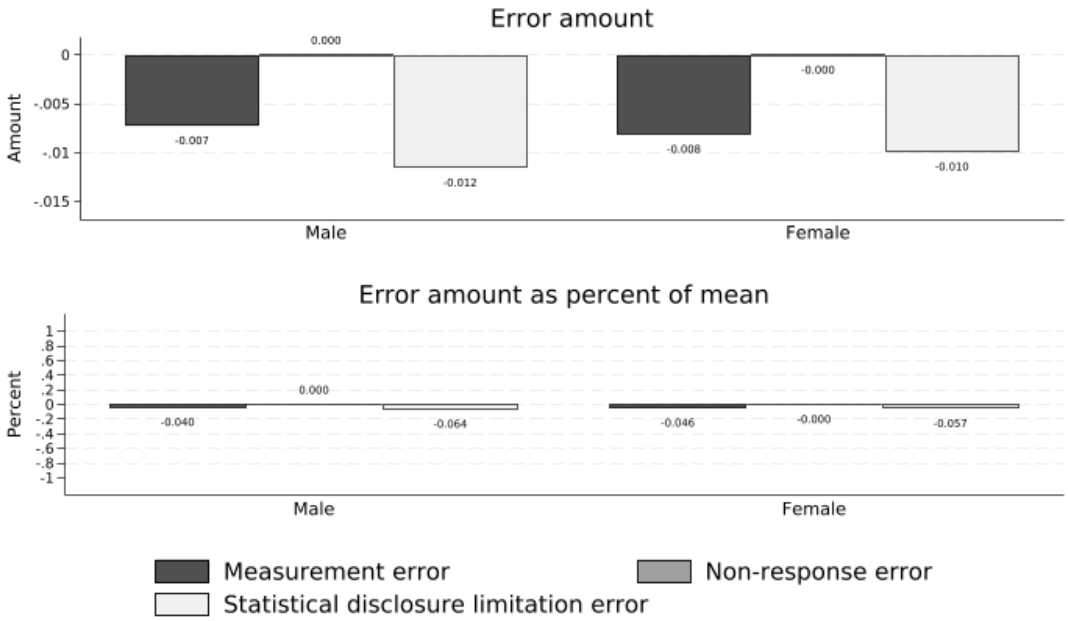
Figure C24: Proportion Hispanic by Education



Source: 2019 American Community Survey and Census Bureau's Best Race and Ethnicity internal file.

Note: The top figure reports error amounts by error type and by education. The bottom figure expresses the amounts from the top figure as a percentage of the survey-based mean proportion Hispanic by education. See Section 2.1 for more information on how the error amounts are computed and Section 4.2 for more information on how error amounts are expressed as a percentage of group-specific means. All results have been reviewed by the U.S. Census Bureau to ensure no confidential information has been disclosed: CBDRB-FY24-CED010-0001.

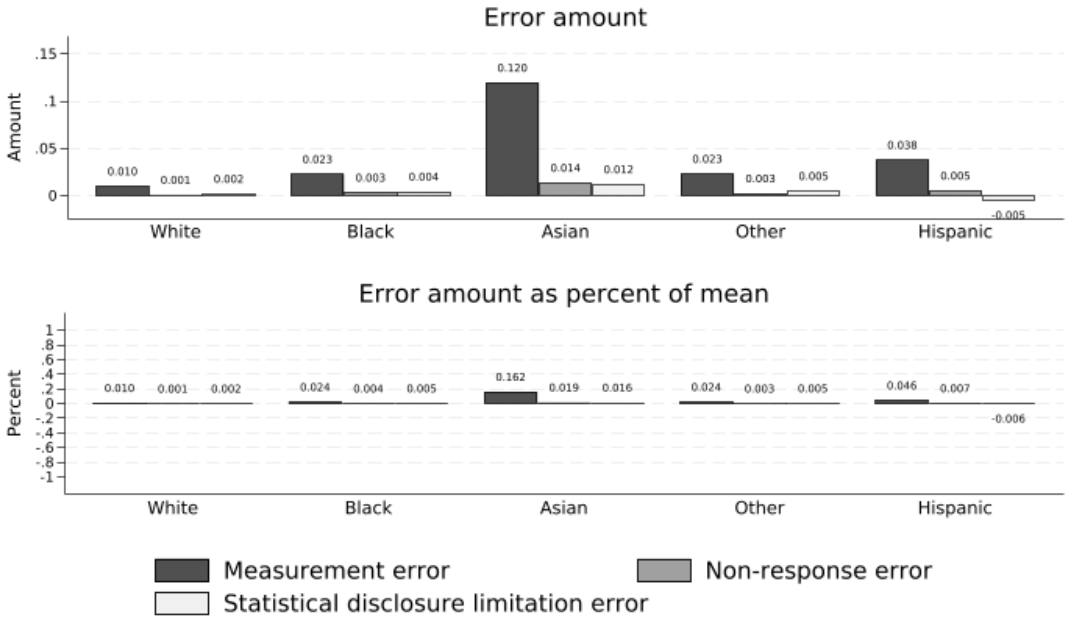
Figure C25: Proportion Hispanic by Sex



Source: 2019 American Community Survey and Census Bureau's Best Race and Ethnicity internal file.

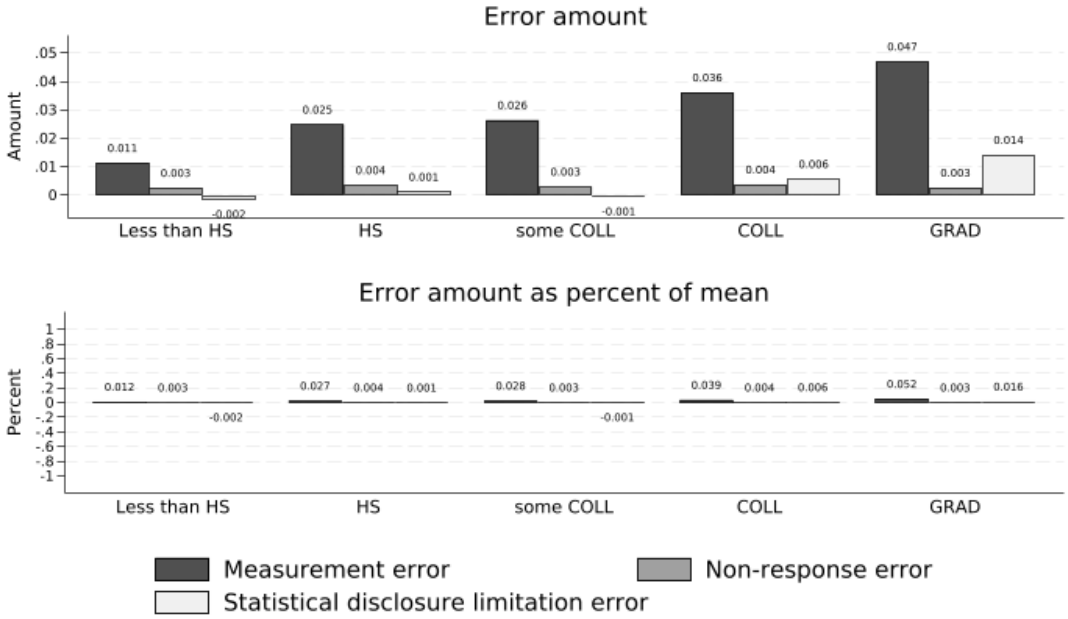
Note: The top figure reports error amounts by error type and by sex. The bottom figure expresses the amounts from the top figure as a percentage of the survey-based mean proportion Hispanic by sex. See Section 2.1 for more information on how the error amounts are computed and Section 4.2 for more information on how error amounts are expressed as a percentage of group-specific means. All results have been reviewed by the U.S. Census Bureau to ensure no confidential information has been disclosed: CBDRB-FY24-CED010-0001.

Figure C26: Proportion Citizen by Race



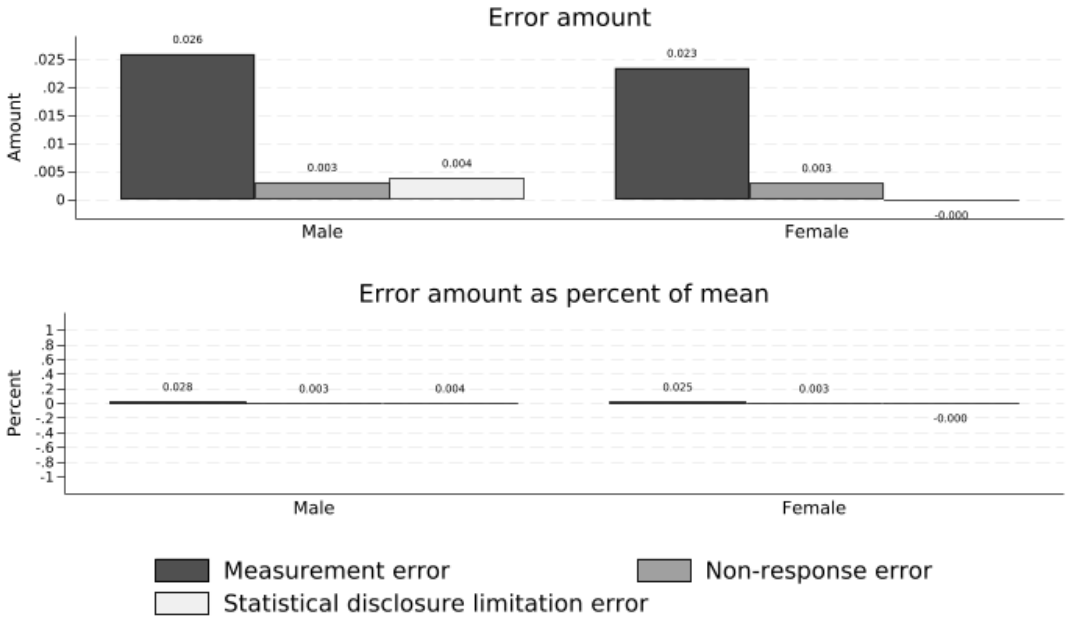
Source: 2019 American Community Survey and Social Security Administration records.
Note: The top figure reports error amounts by error type and by race. The bottom figure expresses the amounts from the top figure as a percentage of the survey-based mean proportion citizen by education. See Section 2.1 for more information on how the error amounts are computed and Section 4.2 for more information on how error amounts are expressed as a percentage of group-specific means. All results have been reviewed by the U.S. Census Bureau to ensure no confidential information has been disclosed: CBDRB-FY24-CED010-0001.

Figure C27: Proportion Citizen by Education



Source: 2019 American Community Survey and Social Security Administration records.
Note: The top figure reports error amounts by error type and by education. The bottom figure expresses the amounts from the top figure as a percentage of the survey-based mean proportion citizen by education. See Section 2.1 for more information on how the error amounts are computed and Section 4.2 for more information on how error amounts are expressed as a percentage of group-specific means. All results have been reviewed by the U.S. Census Bureau to ensure no confidential information has been disclosed: CBDRB-FY24-CED010-0001.

Figure C28: Proportion Citizen by Sex



Source: 2019 American Community Survey and Social Security Administration records.
Note: The top figure reports error amounts by error type and by sex. The bottom figure expresses the amounts from the top figure as a percentage of the survey-based mean proportion citizen by sex. See Section 2.1 for more information on how the error amounts are computed and Section 4.2 for more information on how error amounts are expressed as a percentage of group-specific means. All results have been reviewed by the U.S. Census Bureau to ensure no confidential information has been disclosed: CBDRB-FY24-CED010-0001.