WORKING PAPER · NO. 2019-59

An Empirical Total Survey Error Decomposition Using Data Combination

Bruce D. Meyer and Nikolas Mittag

APRIL 2019

Becker Friedman Institute
FOR ECONOMICS AT UCHICAGO

5757 S. University Ave.
Chicago, IL 60637
Main: 773.702.5599
bfi.uchicago.edu

An Empirical Total Survey Error Decomposition Using Data Combination
Bruce D. Meyer and Nikolas Mittag
April 2019
JEL No. C81,D31,I32,I38

ABSTRACT

Survey error is known to be pervasive and to bias even simple, but important estimates of means, rates, and totals, such as the poverty and the unemployment rate. To summarize and analyze the extent, sources, and consequences of survey error, we define empirical counterparts of key components of the Total Survey Error Framework that can be estimated using data combination. Specifically, we estimate total survey error and decompose it into three high level sources of error: generalized coverage error, item non-response error and measurement error. We further decompose these sources into lower level sources such as a failure to report a positive amount and errors in amounts conditional on reporting a positive value. For error in dollars paid by two large government transfer programs, we use administrative records on the universe of program payments in New York State linked to three major household surveys to estimate the error components we define. We find that total survey error is large and varies in its size and composition, but measurement error is always by far the largest source of error. Our application shows that data combination makes it possible to routinely measure total survey error and its components. The results allow survey producers to assess error reduction strategies and survey users to mitigate the consequences of survey errors or gauge the reliability of their conclusions.

Bruce D. Meyer
Harris School of Public Policy
University of Chicago
1307 E 60th Street
Chicago, IL 60637
and NBER
bdmeyer@uchicago.edu

Nikolas Mittag
Economics Institute of the
Academy of Sciences of
the Czech Republic (CERGE-EI)
Politických vězňů 7
Praha
Czech Republic
and Charles University in Prague
nikolas.mittag@cerge-ei.cz

1. Introduction

Surveys are a key source of information for policy as well as academic research. Survey estimates of means and totals, such as average income, the unemployment rate, as well as the population size of geographic areas or demographic groups, form the basis of key policy decisions. These statistics are also important descriptive information and are used as outcomes or explanatory variables in academic research. Recent studies demonstrate that key household surveys suffer from sizeable, systematic, and growing survey error (Meyer, Mok and Sullivan, 2015). Non-sampling error from sources such as survey coverage, item non-response, and measurement error are often far larger than sampling error. Yet, most studies using survey data only consider how sampling error affects their estimates and fail to account for non-sampling error. One reason for this situation may be the lack of information on error from non-sampling sources and its statistical properties. Survey producers routinely provide users with information on sampling error, but data users are usually left in the dark when it comes to non-sampling error. To understand and address the problem of survey error, we need a framework to measure its extent, its sources, and its likely impact. Once the extent of errors is known, it is possible to devise methods to account for them.

In this paper, we propose to use population data from administrative records or similar sources linked to the survey data to implement a framework to measure non-sampling error systematically and routinely. We define measures of survey error and its components based on the Total Survey Error Framework (Groves and Lyberg, 2010). The total survey error (TSE) of an estimate includes error arising both from sampling and other error sources. It is commonly used at the design stage of surveys to choose between different survey design options. Nevertheless, as Groves and Lyberg (2010) and Biemer (2010) point out in their reviews, TSE is rarely estimated. Similarly, in his discussion of communicating uncertainty in statistics, Manski (2015) concludes that non-sampling error is rarely estimated. The existing estimates often do not provide a general assessment of the multiple sources of error. We argue that data combination, i.e. using information from multiple samples or linking them (see Ridder and Moffitt, 2007), can be used to estimate TSE and to decompose it into components such as generalized coverage error (which combines frame error, unit non-response error and weight adjustments), item non-response error and measurement error. Thereby, data linkage provides a powerful and inexpensive way to analyze and summarize survey error in key statistics such as estimated means and population sizes. Such estimates of TSE can serve as a measure of survey accuracy that helps survey users gauge the reliability of survey estimates. Our decomposition of TSE provides a joint assessment of multiple sources of error in the same survey. The importance of key error components varies between surveys, so that jointly measuring them for each survey is crucial for survey producers to cost-effectively reduce error. The estimated error components can also provide data users with advice on practical questions, such as which survey to use, which variables are reliable, which data problems are most in need of correction, or whether they should use imputed values or not.

Our approach is based on the idea that accurate records on the entire population linked to the survey data can provide the measures of truth required to estimate TSE and its components. We build on results in Meyer, Mok and Sullivan (2015), who compare survey estimates to known population totals to estimate average error in estimated survey means and totals. They also briefly use linked data to decompose aggregate error into components due to item non-response and measurement error as well as a residual component, that combines intentional differences in survey coverage with error arising from sources such as the survey frame and unit non-response. We greatly expand and refine this decomposition by making the known population totals match the population that the survey intends to cover, by introducing corrections for linkage issues, and further decomposing the errors. This reformulation allows us to isolate the survey error in their residual error component, so that it can be interpreted as a measure of generalized coverage error that combines lack of coverage due to survey frame and unit non-response error as well as the effect of weight adjustments. By linking accurate records on the entire population to the survey, we can estimate item non-response error as the difference between survey imputations and the linked, accurate measure, and measurement error as the difference between survey reports and the accurate measure. We split both item non-response and measurement error into two components due to misclassification (i.e. errors in whether the respondent reports zero or not) and a third component due to errors in the continuous part of the variable (i.e. errors in the reported amount if it is not zero). This division is useful for the many important variables that combine a continuous part with a mass point at zero, such as income or expenditures and their components.

We apply this decomposition to average dollars received per household from two government transfer programs, the Supplemental Nutrition Assistance Program (SNAP, formerly the Food Stamp Program) and public assistance (PA), which combines Temporary Assistance for Needy Families and General Assistance. Survey estimates of program participation and dollars received are known to fall short of true receipt (Marquis and Moore, 1990; Taeuber et al., 2004; Lynch et al. 2007; Nicholas and Wiseman, 2009; Kirlin and Wiseman, 2014; Gathright and Crabb, 2015; Meyer, Mok and Sullivan, 2015; Celhay, Meyer and Mittag 2017b; Meyer, Mittag, and Goerge 2018) and this difference is known to affect statistics of importance in policy and academic analyses, such as the income distribution, poverty rates and the effectiveness of the safety net (Bee and Mitchell, 2017; Meyer and Mittag, forthcoming). Therefore, understanding survey error in reported dollars received from government transfers is crucial to assess and improve a key tool of the government to combat poverty and reduce inequality. To implement our decomposition, we use administrative records on the universe of payments from SNAP and PA in New York State (NY) and link them to three large household surveys: the American Community Survey (ACS), the Current Population Survey (CPS) and the Survey of Income and Program Participation (SIPP). These three surveys are the origin of many government statistics and are among the most used household surveys in academic research. Hence, understanding their reliability is important.

We find that TSE is large in all three surveys, ranging from 4 to 60 percent of actual dollars paid. Its size and composition varies both across surveys and across questions within each survey.

Measurement error is by far the largest source of average error and severely understates dollars received from transfer programs. Generalized coverage error and item non-response error are much smaller. They sometimes offset each other, so that without measurement error, TSE would be less than 10 percent of dollars paid in all surveys. Our application shows that TSE could be routinely estimated for major household surveys.

The next section briefly introduces the Total Survey Error Framework and how we define TSE components that can be estimated using data combination. Section 3 introduces our data and describes how we estimate average error and its components for the case of transfer dollars paid. Section 4 presents the results. Section 5 discusses extensions of our data-combination based approach to measure and decompose survey error. Section 6 summarizes our conclusions.

2. Total Survey Error

The Total Survey Error Framework is the dominant paradigm in survey methodology to describe error properties of survey statistics. A common criterion to choose survey design features is to minimize TSE subject to cost considerations (Groves 2004). The TSE of an estimate provides an indicator of survey data quality that goes beyond sampling error by incorporating a variety of error sources. We study average non-sampling error here, because survey users often fail to account for non-sampling error despite the growing evidence that it is a substantial source of error in survey estimates. Numerous theoretical classifications of the sources of non-sampling error exist in the TSE literature, ranging from broad categories such as specification error, frame error, nonresponse error, measurement error and processing error (Biemer and Lyberg, 2003) to detailed classifications including non-statistical measures such as relevance (e.g. Brackstone, 1999; Eurostat 2000; OECD 2003). Groves and Lyberg (2010) provide a detailed discussion of the Total Survey Error Framework and its importance.

As Groves and Lyberg (2010) argue, the success of the Total Survey Error Framework as a theoretical taxonomy to conceptualize the trade-offs in survey design is in stark contrast to the predominant practice of measuring sampling error only. They call the lack of a routine measurement of the full error properties of survey estimates “the great disappointment regarding the total survey error perspective” (Groves and Lyberg, 2010 p. 864). The studies that measure non-sampling error often focus on a specific error source affecting a specific variable in a specific survey, likely due to costs and data availability (Groves and Magilavy, 1986). A few notable exceptions estimate and decompose measures of total error in estimates (Mulry and Spencer 1988, 1991, 1993) and forecasts (Alho and Spencer, 1985) of population sizes. Non-sampling errors in population size can only arise from individuals not being captured by the survey irrespective of their responses to specific questions, so total error in these studies coincides with one of our error components, generalized coverage error.[1] Survey errors vary across surveys and questions (Alwin 2007, Biemer 2009), so it is hard to assess to what extent the available empirical findings generalize. Therefore, it is desirable to develop methods that allow us to jointly study the most important sources of survey error. These methods should be inexpensive, so that they can be routinely implemented.

We argue that data combination can be used to estimate TSE and its key components at a low cost. Databases that contain the entire population of interest are increasingly becoming available from administrative records, private companies, or public sources such as the internet.[2] If these databases can be reliably linked to the survey data, they can provide us with measures of truth for survey concepts, such as the population covered and measures of true responses for key variables. These measures allow us to estimate and decompose TSE. We start with a simple decomposition that loosely follows the approach of Groves et al. (2004) in defining error components corresponding to specific parts of the survey design: generalized coverage error, item non-response error and measurement error. The decomposition could

[1] Another important difference is that these papers decompose error into components arising from different steps and adjustments of complex estimates, whereas our error components capture how different sources of data errors affect (in our case simple) estimates.

[2] We use the terms “population data” and “administrative records” interchangeably for these data below, because we use administrative records in our application.

be extended to other TSE components such as processing error and other measures of TSE, such as mean squared error, as we discuss in section 5.

In particular, we define measures of total non-sampling survey error in the mean of a variable as well as components of this error due to survey coverage, item non-response and measurement error. Our measure of generalized coverage error includes error arising from an incorrect survey frame, unit non-response error and the possibly offsetting effect of any adjustments for these error sources. Survey coverage is of key importance for the estimation of population statistics, the analysis of subpopulations and the allocation of government funds. See de Leeuw and de Heer (2002) and Mulry and Spencer (2001) for surveys of the extent of error and methods to address it. The consequences of (generalized) coverage error in the decennial census of 2000 provide a good example of the importance of understanding and improving survey coverage (Mulry 2007). Item non-response error is the most studied of the three components. Recent validation studies have shown that the common approach to mitigate the consequences of item non-response by imputation leads to high error rates at the household level (Hokayem, Bollinger and Ziliak, 2015; Bollinger et al. 2015; Meyer, Mittag, and Goerge 2018; Celhay, Meyer and Mittag 2017b). These validation studies also provide evidence that measurement error, i.e. respondents providing an inaccurate answer, in variables such as income and education is substantial and pervasive. Bound, Brown and Mathiowetz (2001) provide an extensive review of measurement error in survey data. It is often systematic both in its direction (e.g. underreporting of transfer income in Meyer, Mittag, and Goerge, 2018) and in its relation to other survey variables, which violates the assumptions of common error models and corrections. Past work has not put these errors in a single framework and compared their magnitudes in the same data.

a. Defining a Data Combination Based Measure of Total Survey Error

Formally, we are interested in estimating a parameter μ of a (vector of) random variable(s) X. Let x_i denote the realization of X for unit i. In our application, we estimate TSE in the mean of dollars received from two transfer programs below, so x_i is dollars received by household i and μ is the mean of X. The same formalization can also be used to study TSE in subpopulation means or other parameter. For example, to study survey error in the size of a specific group, x_i would be an indicator of individual i being a member of the subpopulation of interest and μ would be the size of the subpopulation from which x_i is drawn. There are two measures of x_i: x_i^A from the population data, such as our administrative records, which we consider to be accurate and x_i^S from the survey data. Superscripts A and S analogously define the source of other quantities, e.g. μ^A is the population average of x_i^A.[3] Let $\mathcal{P}$ denote the population the survey intends to represent. $\mathcal{S}$ is the population that the actual survey respondents represent, which may differ from $\mathcal{P}$ due to frame error and unit non-response. $\mathcal{L}$ is the set of survey observations that can be linked to the administrative records, which we refer to as the linked data. We assume below that $\mathcal{L}$ can be re-weighted so that it is representative of $\mathcal{S}$. Let $P = |\mathcal{P}|$ be the actual size of the population and P^S the population size assumed by the survey, which may differ from P due to frame error, for example.[4] The survey uses sample or base weights w_i^b that are the inverse of the sampling probability. To account for unit non-response, the survey data may also contain adjusted final weights w_i^f that sum to P^S in the sample of survey respondents. Finally, let r_i indicate whether unit i responded to the relevant survey question.

TSE is the difference between the survey estimate of the parameter of interest and the true value of the parameter for the population the survey intends to cover, μ. We refer to μ as the survey target below. We estimate the TSE as:

$$\hat{\varepsilon}_{TSE} = \frac{1}{\sum_{i\in\mathcal{L}} \widehat{w}_i^{IPW} w_i^f}\left[\sum_{i\in\mathcal{L}} \widehat{w}_i^{IPW} w_i^f x_i^S\right] - \hat{\mu}^A, \qquad (1)$$

[3] We define both μ^A and μ to allow for the possibility of error in the population data, even though we consider them to be accurate in our application. In some cases, μ^A has to be estimated or, as in our case, adjustments have to be estimated. To allow for such cases when the survey target contains an estimated component, we refer to it as $\hat{\mu}^A$ in the formulas below.

[4] We have implicitly simplified the setting to one where P is known rather than estimated and the weights indeed sum to P^S, but generalization to allow for error in the population size or the sum of the weights is straightforward.

where the first term is an estimate of the population mean obtained by weighting the survey reports x^S and the second term is an estimate of the population mean in the administrative data. We estimate the first term from the linked data, using the final weights w_i^f adjusted for incomplete linkage. $\widehat{w}_i^{IPW}$is the inverse probability weight adjustment for incomplete linkage in the survey data discussed in detail below.

If the administrative records cover the same population as the survey and linkage is error free, then it is straightforward to calculate the survey target as the total of x_i^A from the administrative microdata, divided by the population size P. Yet, some of the administrative records may be out of scope for the survey due to intentional differences in coverage (such as the exclusion of long-term care facilities or prisons). In addition, some administrative records may be unlinkable (if, for example, the linking variable is missing), so the linked data cannot cover them. Ideally, we would exclude all records not covered or linkable from the administrative data when calculating the survey target, but the administrative data may not contain the required information to identify and exclude all of these records. If another linkable survey sample covers the subpopulations that are not covered by the linked data, but cannot be excluded from the administrative data, the total of x_i^A for these subpopulations can be estimated and subtracted from the total of x_i^A. Carefully constructing an accurate measure of the survey target is one of the key challenges of measuring TSE. The required subtractions to obtain the survey target as well as the ideal sources of these numbers likely differ between applications. We discuss our case of transfer dollars received in section 3.c. For example, our administrative records include payments to individuals in group quarters that the survey does not intend to cover, as well as payments that are unlinkable, because they were not assigned a PIK. We cannot identify whether a recipient is in group quarters in the administrative data, so we estimate total payments to them using the ACS group quarters sample linked to our administrative records. Our records indicate which observations could not be assigned a PIK, so we subtract total payments on these unlinkable records from total payments in the administrative data. Our application is likely typical in that the subtraction contains both components that we calculate from the administrative data and components that we estimate.

b. Defining Empirical Total Survey Error Components

We first decompose estimated TSE into generalized coverage, item non-response and measurement error. We decompose each of these parts further below. To derive the first decomposition, consider our definition of TSE in equation (1) and replace x_i^S with the equivalent expression $x_i^A + ((1-r_i) + r_i)(x_i^S - x_i^A)$. Multiplying out the parentheses yields the three terms of our decomposition:

$$\hat{\varepsilon}_{TSE} = \frac{1}{\sum_{i\in\mathcal{L}} \widehat{w}_i^{IPW} w_i^f}\left[\sum_{i\in\mathcal{L}} \widehat{w}_i^{IPW} w_i^f x_i^S\right] - \hat{\mu}^A$$

$$= \frac{1}{\sum_{i\in\mathcal{L}} \widehat{w}_i^{IPW} w_i^f}\left[\sum_{i\in\mathcal{L}} \widehat{w}_i^{IPW} w_i^f x_i^A\right] - \hat{\mu}^A + \frac{1}{\sum_{i\in\mathcal{L}} \widehat{w}_i^{IPW} w_i^f}\sum_{i\in\mathcal{L}}(1-r_i)\widehat{w}_i^{IPW} w_i^f\,(x_i^S - x_i^A) +$$

$$\frac{1}{\sum_{i\in\mathcal{L}} \widehat{w}_i^{IPW} w_i^f}\sum_{i\in\mathcal{L}} r_i \widehat{w}_i^{IPW} w_i^f\,(x_i^S - x_i^A)$$

$$= \hat{\varepsilon}_{GCE} + \hat{\varepsilon}_{INR} + \hat{\varepsilon}_{ME}. \tag{2}$$

The first term is generalized coverage error, which is the difference between the error-free estimate of the parameter of interest for the population that the survey actually covers and the survey target:

$$\hat{\varepsilon}_{GCE} = \frac{1}{\sum_{i\in\mathcal{L}} \widehat{w}_i^{IPW} w_i^f}\left[\sum_{i\in\mathcal{L}} \widehat{w}_i^{IPW} w_i^f x_i^A\right] - \hat{\mu}^A\ . \tag{3}$$

If the weights of the survey sample are adjusted to address the problem that the survey sample represents $\mathcal{S}$ rather than $\mathcal{P}$, then this adjustment is included in this term by virtue of using the final weights. Thereby, our estimate of generalized coverage error combines frame error, unit non-response error and the effects of any weight adjustments. It also includes the sampling error of the survey estimate of μ.[5] We discuss more detailed decompositions that further isolate these error sources below.

[5] Adding sampling error in TSE and its components to this framework is conceptually straightforward, but complicates notation. We abstract from sampling variation throughout for two reasons: First, in our large samples this error source is small relative to the bias. Second, sampling variation has already received considerably more attention than the other error components in the prior literature, see e.g. Alwin (1991, 2007) for discussions.

As long as the linked data are representative of the population of interest, data linkage enables us to observe the same variable, x^A for the population and the survey sample. Consequently, we can directly compare (features of) the weighted survey distribution of x^A to its population distribution without any concerns for comparability of the measures in the two data sources. Thereby, data linkage provides us with a simple, but powerful tool to analyze survey coverage. As we discuss in section 5, we do not need x^A to be an accurate measure. Unit non-response can still be analyzed by linking a noisy proxy to both respondents and non-respondents. Survey representativeness can even be analyzed in the case of an uninformative proxy that only indicates whether a unit is present in one or both data sources. We focus on representativeness of the entire survey population, but the same idea can be applied to subpopulations of interest, such as single parent households or the elderly.

Item non-response error is the difference between the statistic of interest calculated using imputed and actual values of x_i for item non-respondents. We estimate item non-response error as:

$$\hat{\varepsilon}_{INR} = \frac{1}{\sum_{i\in\mathcal{L}} \hat{w}_i^{IPW} w_i^f} \sum_{i\in\mathcal{L}} (1-r_i)\cdot \hat{w}_i^{IPW} w_i^f \left(x_i^S - x_i^A\right) \tag{4}$$

We further decompose item non-response error into two parts due to misclassification, false positives (non-recipients who receive benefits according to the survey) and false negatives (recipients not receiving benefits according to the survey), and a part due to errors in amounts among those correctly classified:

$$\hat{\varepsilon}_{INR} = \hat{\varepsilon}_{INR}^{FP} + \hat{\varepsilon}_{INR}^{FN} + \hat{\varepsilon}_{INR}^{Amount} \tag{5}$$

We estimate the three components from the linked data as:

$$\hat{\varepsilon}_{INR}^{FP} = \frac{1}{\sum_{i\in\mathcal{L}} \hat{w}_i^{IPW} w_i^f} \sum_{i\in\mathcal{L}\ s.t\ x_i^S>0\ \&\ x_i^A=0} (1-r_i)\cdot \hat{w}_i^{IPW} w_i^f x_i^S \tag{6}$$

$$\hat{\varepsilon}_{INR}^{FN} = -\frac{1}{\sum_{i\in\mathcal{L}} \hat{w}_i^{IPW} w_i^f} \sum_{i\in\mathcal{L}\ s.t\ x_i^S=0\ \&\ x_i^A>0} (1-r_i)\cdot \hat{w}_i^{IPW} w_i^f x_i^A \tag{7}$$

$$\hat{\varepsilon}_{INR}^{Amount} = \frac{1}{\sum_{i\in\mathcal{L}} \hat{w}_i^{IPW} w_i^f} \sum_{i\in\mathcal{L}\ s.t.\ x_i^S>0\ \&\ x_i^A>0} (1-r_i)\cdot \hat{w}_i^{IPW} w_i^f \left(x_i^S - x_i^A\right) \tag{8}$$

Note that here in the case of decomposing a mean the terms for false positives and false negatives will at least partially offset each other. In the case of other statistics such as mean squared error, they would generally not offset. Finally, measurement error is the difference between the statistic of interest calculated using reported and actual values of x_i for respondents. Estimating this component from the linked data is a straightforward modification of how we estimate item non-response error:

$$\hat{\varepsilon}_{ME} = \frac{1}{\sum_{i\in\mathcal{L}} \hat{w}_i^{IPW} w_i^f} \sum_{i\in\mathcal{L}} r_i \cdot \hat{w}_i^{IPW} w_i^f \left(x_i^S - x_i^A\right) \tag{9}$$

As above, we decompose measurement error into parts due to false positives, false negatives and errors in amounts:

$$\hat{\varepsilon}_{ME}^{FP} = \frac{1}{\sum_{i\in\mathcal{L}} \hat{w}_i^{IPW} w_i^f} \sum_{i\in\mathcal{L}\ s.t\ x_i^S>0\ \&\ x_i^A=0} r_i \cdot \hat{w}_i^{IPW} w_i^f x_i^S \tag{10}$$

$$\hat{\varepsilon}_{ME}^{FN} = -\frac{1}{\sum_{i\in\mathcal{L}} \hat{w}_i^{IPW} w_i^f} \sum_{i\in\mathcal{L}\ s.t\ x_i^S=0\ \&\ x_i^A>0} r_i \cdot \hat{w}_i^{IPW} w_i^f x_i^A \tag{11}$$

$$\hat{\varepsilon}_{ME}^{Amount} = \frac{1}{\sum_{i\in\mathcal{L}} \hat{w}_i^{IPW} w_i^f} \sum_{i\in\mathcal{L}\ s.t.\ x_i^S>0\ \&\ x_i^A>0} r_i \cdot \hat{w}_i^{IPW} w_i^f \left(x_i^S - x_i^A\right) \tag{12}$$

With additional information, more detailed decompositions can be implemented. For example, if the base weights (i.e. weights before any nonresponse correction), w^b, are available, one can first replace w^f in equation (1) by $w^b + (w^f - w^b)$:

$$\hat{\varepsilon}_{TSE} = \frac{1}{\sum_{i\in\mathcal{L}} \hat{w}_i^{IPW} w_i^f} \left[\sum_{i\in\mathcal{L}} \hat{w}_i^{IPW} w_i^b x_i^S + \sum_{i\in\mathcal{L}} \hat{w}_i^{IPW} (w^f - w^b) x_i^S\right] - \hat{\mu}^A \tag{13}$$

Replacing x_i^S in the first term with $x_i^A + ((1 - r_i) + r_i)\left(x_i^S - x_i^A\right)$ as in equation (2) yields a decomposition that additionally isolates the effect of weight adjustments on survey error:

$$\begin{aligned}
\hat{\varepsilon}_{TSE} &= \frac{1}{\sum_{i\in\mathcal{L}} \hat{w}_i^{IPW} w_i^f} \left[\sum_{i\in\mathcal{L}} \hat{w}_i^{IPW} w_i^b x_i^A + \sum_{i\in\mathcal{L}} \hat{w}_i^{IPW} \left(w^f - w^b\right) x_i^S\right] - \hat{\mu}^A \\
&\quad + \frac{1}{\sum_{i\in\mathcal{L}} \hat{w}_i^{IPW} w_i^f} \left[\sum_{i\in\mathcal{L}} (1 - r_i) \hat{w}_i^{IPW} w_i^b \left(x_i^S - x_i^A\right) + \sum_{i\in\mathcal{L}} r_i \hat{w}_i^{IPW} w_i^b \left(x_i^S - x_i^A\right)\right] \\
&= \hat{\varepsilon}'_{GCE} + \hat{\varepsilon}'_{INR} + \hat{\varepsilon}'_{ME}.
\end{aligned} \tag{14}$$

The main difference from equation (3) is that generalized coverage error now includes an additional term:

$$\hat{\varepsilon}'_{GCE} = \frac{1}{\sum_{i\in\mathcal{L}} \hat{w}_i^{IPW} w_i^f} \left[\sum_{i\in\mathcal{L}} \hat{w}_i^{IPW} w_i^b x_i^A + \sum_{i\in\mathcal{L}} \hat{w}_i^{IPW} (w_i^f - w_i^b) x_i^S\right] - \hat{\mu}^A \tag{15}$$

The first sum in the brackets is the weighted total of the administrative variable in the linked data. It uses the base weights, so that it provides an estimate of the total of x^A among the population of survey respondents, $\mathcal{S}$. The second sum in brackets is the effect of the weight adjustment.[6]

In this alternate decomposition, generalized coverage error includes the actual weight adjustment using survey reports x_i^S rather than the adjustment using x_i^A, which is implicitly included in $\hat{\varepsilon}_{GCE}$ above in the earlier version of the decomposition. The difference, $\frac{1}{\sum_{i\in\mathcal{L}} \widehat{w}_i^{IPW} w_i^f} \sum_{i\in\mathcal{L}} \widehat{w}_i^{IPW} (w_i^f - w_i^b)(x_i^S - x^A)$, arises from both coverage and measurement issues. Thus, one can include the parts of this interaction term arising from item non-response and measurement error in their respective components instead, as we did in the original decomposition. For item non-response and measurement error, subtracting this interaction term is the only difference between our two decompositions. Therefore, the expressions for these two error components use the base weights in equation (14) instead of the final weights as in equation (2). An advantage of making this interaction term part of generalized coverage error is that any error in the actual coverage adjustment is attributed to coverage error. At the same time, it makes our weighted sample of those with item non-response error and measurement error representative of the population of item non-respondents and respondents. Thus, statistics such as average dollars received from our weighted sample are unbiased estimates of the corresponding parameters for the population of item non-respondents and respondents.

These components can be decomposed further. The effect of the weight adjustment is straightforward to estimate if both final and base weights are available. One can split the remainder of generalized coverage error, $\frac{1}{\sum_{i\in\mathcal{L}} \widehat{w}_i^{IPW} w_i^f} \sum_{i\in\mathcal{L}} \widehat{w}_i^{IPW} w_i^b x_i^A - \hat{\mu}^A$ into frame error and unit non-response

[6] Thus, one can also decompose TSE into four components by separating error due to weight adjustments from ε'_{GCE}, i.e. by defining $\hat{\varepsilon}''_{GCE} = \frac{1}{\sum_{i\in\mathcal{L}} \widehat{w}_i^{IPW} w_i^f} \sum_{i\in\mathcal{L}} \widehat{w}_i^{IPW} w_i^b x_i^A - \hat{\mu}^A$ and $\varepsilon''_{WGT} = \frac{1}{\sum_{i\in\mathcal{L}} \widehat{w}_i^{IPW} w_i^f} \sum_{i\in\mathcal{L}} \widehat{w}_i^{IPW} (w_i^f - w_i^b) x_i^S$.

error by linking population data to the sampling frame rather than only sampled units as we discuss in Section 5.

c. Assumptions

We assume throughout that using the linked sample with weights adjusted for survey households that cannot be linked is (asymptotically) equivalent to using the full survey sample. To adjust the survey weights, we multiply them by $\widehat{w}_i^{IPW} = \widehat{\Pr}(PIK_i = 1|Z_i)^{-1}$ where PIK_i is an indicator for whether an identifier for linkage (a PIK) was obtained for survey unit i, i.e. whether the unit can be linked, and Z_i is survey variables used in a model of the probability that survey household i can be linked.[7] If missing PIKs are independent of X conditional on Z, this reweighting makes the linked data representative of the survey population. Alternatively, one could estimate the survey mean from the entire survey sample that includes the unlinked observations. Using the linked data average is natural, as it allows us to keep the sample fixed in the decomposition, thus making the error components sum to $\hat{\varepsilon}_{TSE}$. Under the (testable) assumption that linkage status does not affect the conditional expectation of x_i^S given Z_i, i.e. $\mathbb{E}(x_i^S | PIK_i = 1, Z_i) = \mathbb{E}(x_i^S | Z_i)$, both estimates from the entire survey sample and the linked sample consistently estimate the mean of x^S. We neither reject this assumption in our data nor do the two estimates differ by meaningful amounts. In applications with lower match rates or larger differences between the estimates, the survey mean is likely preferable to the mean from the linked sample.

A related result is that our estimate of generalized goverage error is consistent as long as conditional on Z, the expectation of x^A does not depend on linkage status, i.e. $\mathbb{E}(x_i^A | PIK_i = 1, Z_i) = \mathbb{E}(x_i^A | Z_i)$.[8] Consistency of $\hat{\varepsilon}_{INR}$ and $\hat{\varepsilon}_{ME}$ require this assumption to hold among item non-respondents and respondents, i.e. $\mathbb{E}(x_i^A | PIK_i = 1, Z_i, r_i = 0) = \mathbb{E}(x_i^A | Z_i, r_i = 0)$ and $\mathbb{E}(x_i^A | PIK_i = 1, Z_i, r_i = 1) =$

[7] See section 3.b and Appendix 2 for descriptions of how we implement this adjustment in our application.

[8] Contrary to the analogous assumption on $\mathbb{E}(x_i^S | Z_i)$ above, we do not observe x^A for observations that cannot be linked. Some information on its validity can be obtained by examining whether linkage status predicts related variables, such as survey reports x^S.

$\mathbb{E}\left(x_i^A \middle| Z_i, r_i = 1\right)$. Similarly, the decomposition into error due to false positives, false negatives and errors in amounts requires the assumption to hold among each of the three subpopulations.

3. Application: Using Linked Data to Decompose Total Survey Error in Program Receipt

We decompose the error in mean dollars received per household from two government transfer programs, the Supplemental Nutrition Assistance Program (SNAP, formerly the Food Stamp Program) and public assistance (PA), which combines Temporary Assistance for Needy Families and General Assistance. Understanding survey error in reported dollars received from these government transfers is crucial when assessing government efforts to combat poverty and reduce inequality.

a. Data Sources

Our population data are administrative records from the New York State Office of Temporary and Disability Assistance (OTDA) for all SNAP and PA recipients in the state. The records include monthly payment amounts and dates from 2007 through 2012, as well as basic demographic information and addresses. The accuracy of the individual identifiers and amounts paid is crucial to the validity of our estimates and they appear to be of high quality. As part of eligibility determination, applicant information is checked by OTDA against social security records. The data are from actual payments and are audited. The overall total from our administrative records matches official aggregate reports by OTDA almost exactly.[9] This agreement provides evidence of the accuracy of our administrative microdata and justifies our use of total amounts from our administrative microdata to estimate the survey target below.

We study TSE in dollars received from SNAP and PA in the ACS, the CPS, and the SIPP. Appendix 1 and the survey documentation (U.S. Census Bureau 2006, 2008, 2014) provide detailed information on the surveys and their design. The ACS is the largest household survey in the U.S., with more than 290,000 households selected each month to participate. The CPS is one of the most important economic surveys

[9] We are only able to make this comparison for the SNAP records as published aggregates comparable to our PA administrative data are not available.

in the U.S. with 60,000 households participating in the survey each month of the year. It is the official source of labor force statistics. We use the Annual Social and Economic Supplement (ASEC) of the CPS, which also is the official source of income and poverty statistics in the U.S. Finally, the SIPP is the highest quality source of information on low income households and the receipt of government transfers. The 2004 and 2008 SIPP panels that we use initially sampled approximately 50,000 households and intended to follow them for a period of 4 years.

We use the sample of households in NY State from all three surveys. The ACS and the CPS are representative at the state level. The SIPP is representative at the national level, but not claimed to be representative at the state level.[10] All surveys cover the entire population residing in households, which is our population of interest. The ACS also includes a group quarters sample that is described in U.S. Census Bureau (2014, chapter 8). It covers the population living in residential structures that are not households, which includes those living in college residence halls, residential treatment centers, group homes, military barracks, correctional facilities, and dormitories, among others. Thereby, the ACS is representative of the entire residential population in the US, so that only individuals who do not live in residential structures are neither included in our samples of interest, nor in the ACS group quarters sample.

The three surveys are large-scale, general interest surveys, which makes them similar in survey design. Still, there are also pronounced differences in survey design features known to be related to both non-response and measurement error, see Celhay, Meyer and, Mittag (2017a) for a discussion. All three surveys use similar sample frames based on (augmented) extracts from a master address file. The surveys take unit non-response into account by assigning weights based on the household sampling probabilities to all individuals in the household and then adjust these individual base weights to make demographic

[10] Thus, our estimate of ε_{REP} for the SIPP is an estimate of coverage differences for NY State that may not arise from coverage error. Our estimates of item non-response and measurement error are based on household level errors, so they still provide estimates of the respective error for the population covered by the NY SIPP sample.

characteristics of the sample match population statistics. Due to the complex nature of the weights,[11] we use the final weights and do not implement the more detailed decomposition in equation (14).

The ACS and CPS only interview one household member, but the SIPP strives to conduct in-person interviews with every member of the household over age 15 every four months. In terms of information on government transfers, all three surveys ask for receipt of SNAP and PA. The surveys also collect information on amounts received, but the ACS does not ask for amounts received from SNAP. Consequently, for SNAP in the ACS we are only able to estimate generalized coverage error. For both programs, the questions in the ACS refer to the 12 months prior to the interview date. The CPS asks about the previous calendar year and the SIPP asks for monthly information in the four months before the interview month. Thus, the time periods covered by our three survey samples differ. Our ACS sample covers 2008-2012, the CPS sample covers calendar years 2007-2012 and the SIPP sample covers 2007 (wave 10-12 of the 2004 SIPP panel) and August 2008 to December 2012 (wave 1-14 of the 2008 SIPP panel). For comparability, we report annual averages throughout.[12]

All three surveys impute missing values due to item non-response using a hot deck procedure. The CPS and SIPP also impute multiple items or entire records from a single donor in some cases.[13] See U.S. Census Bureau (2006, 2008, 2014) for detail and Meyer, Mittag, and Goerge (2018) for a summary. Our imputed sample includes both cases where only the program receipt or amount question is imputed and cases where multiple items were imputed from a single donor. For the SIPP, we consider a (monthly)

[11] See Appendix 1 for an overview of the weight adjustments in the three surveys we examine.

[12] Reference periods differ between households in the ACS, which makes estimates for each of our five survey years a weighted average of a 24-month period. See Chapter 7 in U.S. Census Bureau (2014) for details, we account for this sampling scheme by calculating all payments based on weighted monthly or annual numbers. Using the intended, the actual or a uniform distribution of interviews across months to weight monthly numbers does not affect our results in a meaningful way. The annual SIPP estimates in our analysis are estimated using monthly total payments and then aggregated to the annual level as suggested by the SIPP documentation. Imputation and error rates are in terms of the unit of observation, i.e. household-months for the SIPP and household-years for the CPS and ACS.

[13] The CPS imputes the entire record for those who answer the CPS, but not the ASEC (the so-called whole imputes). The SIPP mainly imputes entire records for non-interviewed persons within interviewed households.

response to be imputed if it was imputed for any household member. The resulting imputation rates differ substantially between surveys and questions. For SNAP, our imputed sample accounts for 9.0 percent of the population in the CPS, and 7.8 percent in the SIPP. For PA, it accounts for 6.1 percent of the population in the ACS, 3.2 percent in the CPS, and 7.6 percent in the SIPP.

b. Data Linkage

We link the administrative data to the three surveys at the household level using person identifiers created by the Person Identification Validation System (PVS) of the U.S. Census Bureau.[14] Celhay, Meyer and Mittag (2017a,b) and Meyer, Mok, and Sullivan (2015) use the same linked data and further discuss data linkage and accuracy. The PVS was applied to make the administrative data and the entire U.S. survey samples linkable, including the ACS group quarters sample. In short, the PVS uses the person data (such as address, name, gender, and date of birth) from the administrative records and survey data to search for a matching record in a reference file that contains all transactions recorded against a social security number. If a matching record is found, the social security number of the record from the reference file is transformed into a protected identification key (PIK)[15] and attached to the corresponding records in our data. For the administrative records, a PIK is obtained for over 99 percent of the records from each program. We can link the information from a program case to the correct survey household if any true recipient member is assigned a PIK.[16] Therefore, we consider a household to have a PIK if a PIK was obtained for someone in the household. The PIK rates at the household level are 93, 91, and 95 percent in the ACS, CPS and SIPP, respectively.

[14] NORC (2011) and Wagner and Layne (2014) discuss the PVS in detail.

[15] PIKs are anonymized social security numbers and used to protect the identity of individuals in the data.

[16] The administrative records contain every individual on the case, so one PIKed household member is sufficient for us to match receipt correctly except for households in which all PIKed members are true non-recipients, but there are true recipients among the non-PIKed members. Usually only a few PIKs are missing per household (89 percent of individuals are PIKed in the ACS and 86 percent in the CPS and SIPP) and few non-recipients cohabit with recipients, so these exceptions should be uncommon.

Matching both the survey and the administrative data to a third data source has the advantage that we can distinguish between unlinkable records and linkable records that are not in the other data source, because the former do not have a PIK. This setting allows us to adjust for incomplete linkage in both data sources. We account for unlinkable administrative records by subtracting total payments without a PIK from the survey target as discussed below. We cannot validate receipt information for survey households without a PIK, so our analyses are based on the PIKed survey sample (the "linked data"). Despite the low rate of missing PIKs, they are not missing completely at random in the survey data. To restore representativeness of the linked data for the NY household population, we use inverse probability weighting (IPW, see Wooldridge 2007). To do so, we estimate Probit models to predict the probability that a household has a PIK and multiply the survey weights by the inverse of this predicted probability.[17] This adjustment relies on the assumption that conditional on the covariates in the Probit model, whether a household has a PIK or not does not predict transfer dollars paid or reported. For dollars reported, this assumption is testable. In our samples, it holds for both SNAP and PA. For PA, it holds unconditionally. Even before reweighting, the difference between the sample with PIKs and the entire survey sample is less than 5 dollars per household in all cases. As the high rate of PIK-linking suggests, our results do not appreciably change when using the unadjusted household weights. To assess survey coverage of subpopulations, one needs to make sure to include subpopulation status as a covariate in these Probit models. Otherwise the re-weighted data may be representative of the overall population, but not the subpopulation of interest.

c. Estimating Empirical Survey Error Components

A key challenge in measuring TSE is estimating the survey target, $\hat{\mu}^A$ as discussed in section 2.a. We first calculate total program dollars paid to the entire population from our administrative microdata. The

[17] We estimate separate Probit models to adjust weights in the group quarters file and for households outside of NY. See Appendix 2 for further detail.

records indicate the day of the payment, so we can exactly match the time period of each survey sample. The administrative microdata include both payments that the survey data do not intend to cover and payments that cannot be linked at all or cannot be linked to our NY sample. To make our survey target comparable to the linked data, we need to subtract these payments from the total according to the entire administrative microdata.

In terms of intentional differences in coverage, the survey samples we use cover the U.S. household population, but do not include those living in group quarters and non-residential structures. Consequently, to make the administrative total comparable to the population our survey samples intend to cover, we need to estimate total dollars paid to these two groups. Our administrative records include the street address of the recipient, so we can identify payments to individuals without a residential address.[18] This information allows us to calculate total payments to the population in non-residential structures from the administrative records. Whether a recipient lives in group quarters is not recorded in the administrative microdata, so subtracting these payments from the survey target requires estimating them from another source. We estimate total amounts paid to individuals in group quarters from the 2008-2010[19] ACS group quarters file linked to our NY administrative records. Subtracting these two estimates from total dollars paid according to the administrative records leaves us with a measure of total SNAP and PA amounts paid to the New York population residing in households.

Differences between the population data and the linked data may also arise from the linkage process. We account for two differences arising from data linkage: total dollars according to unlinkable

[18] We use a string match of the address field with common terms for “undomiciled” and “homeless” to identify additional non-residential individuals, most of whom are street homeless. For some of them, the address field may be missing or contain an address for mail only, so our approach likely slightly underestimates payments to non-residential individuals.

[19] The change in the group quarters sample in 2011 (U.S. Census Bureau, 2012) makes estimating comparable state-level totals difficult. Thus, we only use the data for 2008-2010 and extrapolate to match the time periods covered by our surveys based on the assumption that the fraction of total program dollars paid to group quarters is constant over time. This assumption fits the data over 2008-2010 better than an assumption of a constant dollar amount paid to those in group quarters.

administrative records, and links to households outside of NY State. Under the assumption given earlier, the inverse probability weight adjustment corrects for survey units that cannot be linked. There are often also administrative records that are unlinkable, because no PIK was obtained for them, so they are not covered by the linked data. We combine the administrative data and the survey data by linking each data set to a third data source, so we can distinguish unlinkable administrative records from those that did not receive the program, because the former do not have a PIK. This information allows us to calculate the total amount paid to unlinkable records as the total amount according to records without PIKs in the administrative data and account for this difference when calculating the survey target.[20]

In addition, some of our administrative records are linked to households that are in the survey, but excluded from our analysis. We only study NY state here, so our linked data do not capture payments from NY OTDA to individuals residing outside of NY at the time of the survey. Individuals are only allowed to receive SNAP and PA in the state they reside in, but this occurrence may be due to mobility during the reference period of the survey, having a second residence or temporarily working or serving in the army in another state.[21] We link the administrative records from NY to the entire U.S. survey samples, so we can estimate total NY SNAP and PA payments to individuals residing in other states at the time of the survey and subtract it from the administrative total.

Consequently, our final estimate of the target for each survey is total dollars paid according to the administrative records minus dollars paid to the population in non-residential structures, estimated

[20] While the administrative records without a PIK cannot be captured by x^A, they may be reported and thus included in x^S. This situation would make our approach slightly understate the extent of underreporting. Alternatively, one could add the amounts to the estimated survey total. The alternative relies on the assumption that the survey would capture these missing payments if it could. This approach amounts to adding the same small number to the numerator and denominator of our estimates and would only have a negligible effect in our application.

[21] Payments to households in other states may also arise from linking records to the wrong PIK. For the analyses here, it is not important whether these payments are linkage errors or true receipt by residents of other state: in both cases, we should not expect the linked NY sample to capture them. Further analyses of these records suggest that linkage errors are likely at most infrequent: For example, almost 70 percent of PIKs linked to non-NY households report either moving or serving in the military. It is also re-assuring that most of these households include multiple recipients of NY payments, which would be unlikely with random linkage errors.

dollars paid to those in group quarters, dollars paid according to unlinkable administrative records, and estimated dollars paid to those residing outside of NY state, divided by the number of households in NY.[22] As defined above, our estimate of TSE is the difference between average reported dollars estimated from the re-weighted linked data, and this survey target.

We decompose estimated TSE into three components as defined by equation (2): generalized coverage error, item non-response error and measurement error.[23] We estimate generalized coverage error as the survey estimate of average dollars paid according to the linked administrative variable minus the estimated survey target $\hat{\mu}^A$. To estimate item non-response and measurement error, we calculate survey error for each linked household. For item non-respondents, the error is the difference between their imputed amount and the linked amount from the administrative records. For respondents, household level survey error is the difference between the reported amount and the linked administrative amount. As defined in section 2.a, we estimate item non-response error $\hat{\varepsilon}_{IRN}$ and measurement error, $\hat{\varepsilon}_{ME}$, as the estimated population total of this difference among item non-respondents and respondents divided by the number of households P. We further decompose our estimates of item non-response and measurement error into two parts due to misclassification and a part due to errors in amounts. We estimate net misclassification error as the difference between error due to false positives (total dollars reported by non-recipients according to the linked data) and error due to false negatives (total dollars received by those who do not report receipt according to the administrative variable) divided by P. Our

[22] Constructing an independent estimate of the size of the household population based on Census figures does not make a meaningful difference to our estimates for the ACS and SIPP, so we use the population size implied by the survey weights, i.e. $P = P^S = \sum_{i \in \mathcal{L}} \widehat{w}_i^{IPW} w_i^f$. The CPS weights overstate the number of households as discussed in section 3.a. Thus, instead of using P^S as implied by the weights, we use the population size implied by the ACS for the corresponding time period as P. Yet, we do not include the difference due to the higher aggregate weights, $\left[\left(\sum_{i \in \mathcal{L}} \widehat{w}_i^{IPW} w_i^f\right)^{-1} - P^{-1}\right]\left[\sum_{i \in \mathcal{L}} \widehat{w}_i^{IPW} w_i^f x_i^S\right]$, in our estimate of TSE. Doing so would increase our estimate of TSE, but this error component is not present in individual level statistics and seems simple for survey users to address. Note that this convention only affects the estimates reported in terms of dollars in Table 2. The population size cancels when reporting error components as a share of TSE, which we do throughout.

[23] As discussed in section 3.a, the weight adjustments are complex, and the base weights are not available in our data. Therefore, we do not implement the more detailed decomposition defined by equation (14).

estimate of error in amounts is the estimated population total of household-level survey error among households recorded as receiving the program according to both the survey and the administrative variable divided by P.

4. Results

a. Total Survey Error

Table 1 summarizes our subtractions to obtain the survey target from the official total outlays. The overall subtractions are of a similar magnitude across surveys, but slightly higher for PA (12-17 percent) than for SNAP (8-10 percent). For all surveys and both programs, the payments to group quarters are the largest component and make up more than 50 percent of the overall subtraction. This size makes the subtraction for intentional difference in coverage account for around two-thirds of the total subtraction. The subtraction for unlinkable administrative records is less than two percent of the overall amount paid in all cases. Subtracting these payments from the administrative data total yields the total in the last row of Table 1. For both SNAP and PA, the amounts are similar across surveys, but vary slightly due to the differences in the years covered.

Table 1: Total SNAP and PA Payments that Should be Covered by the Linked NY Household Sample in Three Surveys (AROUND HERE)

Table 2: TSE in Average Annual SNAP and PA Dollars Paid to NY Households for Three Surveys (AROUND HERE)

Dividing the total amount in the last row of Table 1 by the number of households in the corresponding population yields the survey target in the first row of Table 2: average dollars paid per household. TSE is the difference between average dollars reported in the linked data (provided in the next row of Table 2) and this number. As the last two rows of Table 2 show, all surveys fall short of their target. TSE is substantial both in absolute and relative terms, in all cases except for SNAP in the SIPP. It ranges from missing 4.3 percent of average dollars (SNAP in the SIPP) to missing almost three out of five dollars of PA in the CPS. The difference between programs is also large: the fraction of dollars missed is much higher for PA, so that the surveys fail to capture a similar amount for PA and SNAP, despite SNAP being more than twice as large as PA. Across surveys, TSE is remarkably similar for PA. For SNAP, we only have estimates for the CPS and the SIPP, because the ACS does not ask for SNAP amounts. The SIPP misses only 4.3 percent of SNAP dollars, which is remarkable in comparison to the 30 percent of SNAP dollars missing in the CPS and the large PA amounts missing in all surveys. The comparability of the population covered by the SIPP and the administrative data is questionable, so this finding should be interpreted with caution. Consequently, our results are not conclusive regarding which survey does best. Rather, they can be taken as evidence that which survey is most accurate varies even for similar questions.

b. Decomposing Total Survey Error

Figure 1 and Table 3 present the results of our TSE decomposition. The size of the vertical bars in Figure 1 indicates survey error as a fraction of the survey target. For convenience, Table 3 also provides these numbers for each error component. The numbers in each error component of Figure 1 provide the error due to each component as a percentage of TSE for the respective survey and program.

Figure 1: TSE Components in Percent of Average Dollars Paid and as Shares of TSE (AROUND HERE)

Table 3: Total Survey Error and its Components as Share of Average Dollars Paid per Household (AROUND HERE)

A first thing to note is that while aggregate TSE is negative in all cases, some error components are positive. Generalized coverage error in the CPS and the SIPP for both programs as well as item non-response error for PA in the CPS offset other sources or error. The degree to which two wrongs make a right, because some components reduce overall error, varies across surveys. As the case of SNAP shows, offsetting errors can make aggregate error understate the differences in the extent of errors across surveys. The large difference between the CPS and SIPP is partly due to the much larger over-representation of dollars paid in the SIPP. The case of PA shows that offsetting errors can also have the opposite effect and understate differences across surveys. While all three surveys are similar in terms of TSE for PA, the decomposition shows that there is much less error in the ACS than in the CPS, which in turn has slightly less error than the SIPP. These differences underline the importance of not only measuring TSE, but also decomposing it into its components. While offsetting errors improve net dollars in the survey, they may increase bias for subpopulations and estimates of model parameters.

The sign and the magnitude of the error components vary across surveys and programs. The only component that leads surveys to understate dollars paid in all cases is measurement error. While it is always large, the measurement error bias varies across surveys and questions as well. Together with the variation in the other components, this makes the composition of TSE differ substantially across the columns of Figure 1. This finding further emphasizes that measures of net survey error can conceal differences between surveys and questions not only in the extent of error, but also in its sources.

Generalized coverage error ranges from -4.5 to 10.7 percent of average dollars paid. Thereby, it is substantially smaller than measurement error, but larger than item non-response error in all cases. Contrary to the other two components, generalized coverage error is similar for SNAP and PA within the same survey. This is not surprising, because the population of program recipients is similar, so one would expect a survey that covers recipients of one program well to also cover recipients of the other program well. Between surveys, generalized coverage error varies both in terms of sign and magnitude. Generalized coverage error in the ACS is negative for both programs, but minimal at -1 percent for SNAP and only slightly larger for PA at -4.5 percent. For the CPS and the SIPP, our estimates suggest that recipients of both programs are overrepresented in the NY sample of the two surveys. At slightly above 5 and 6 percent for SNAP and PA respectively, the overrepresentation of receipt in the CPS is relatively small for both programs. Households represented in the NY SIPP sample on average receive about 10 percent more from both programs than households in NY actually receive. Since the SIPP is not claimed to be representative of individual states, these numbers could be interpreted as encouraging evidence that the SIPP is informative about state-level program receipt.

Item non-response error is the smallest component of TSE in all cases. Apart from PA in the CPS, the error from item non-response is negative, i.e. the surveys slightly under-impute dollars received. For all surveys and programs, the difference between imputations and actual receipt by item non-respondents is less than 3.1 percent of average dollars received. These numbers show that the high imputation error prior studies found at the household level (Meyer, Mittag, and Goerge 2018, Celhay, Meyer, Mittag 2017b) by and large cancels in the (unconditional) aggregate. Therefore, while the prevalence of error at the household level makes the use of imputed observations in economic models or studies of subpopulations questionable (Hirsch and Schumacher 2004; Bollinger and Hirsch 2006; Celhay, Meyer and Mittag 2017b), for our surveys and programs, imputation comes close to achieving its key objective of correcting overall averages.

As Figure 1 shows, measurement error is by far the largest error component for both programs in all three surveys. The error due to misreporting among respondents is always larger than the other two error components combined. With the exception of SNAP in the SIPP, measurement error is 6 to 30 times larger in absolute value than each of the other two components. The large extent of measurement error and that the other error components partly offset each other implies that measurement error accounts for almost the entire net understatement of dollars received in all cases. Without measurement error, the surveys would err by -6.6 to 9.5 percent rather than understating dollars received by 4.3 to 59.8 percent.

Measurement error is large and negative in all cases, but the results also suggest systematic differences across surveys: The CPS misses the largest amount for both programs. At the other extreme, net underreporting is only 13.9 percent for SNAP in the SIPP on average. In the case of PA, survey responses miss 65.8 (SIPP) and 68.5 (CPS) percent of PA dollars paid. Net underreporting is slightly lower in the SIPP than in the CPS, but alarmingly high in both surveys. In a similar vein, PA dollars reported in the ACS are more accurate than the SIPP, but the ACS still misses an impressive 50.5 percent of dollars paid due to measurement error alone. Underreporting of PA is more severe than underreporting of SNAP, which accounts for a large fraction of the difference in TSE between SNAP and PA.

We further decompose item non-response and measurement error into a component arising from errors in receipt status, which in turn consists of errors due to false positives and false negatives, and a component due to errors in amounts when receipt status is correct. As pointed out above, many other variables have such a mixed continuous nature. We often study these intensive and extensive margins separately, e.g. studies of work effort or transfer receipt and amounts received. Consequently, understanding from which margin the error stems is important to gauge the accuracy of such studies.

Table 4: Sources of Item Non-Response Error and Measurement Error as Shares of Average Dollars Paid (AROUND HERE)

For item non-response, most surveys also use separate imputation procedures for the binary and the continuous part of such variables. Thus, improving imputations requires understanding which of the two components drives the error. The first row of Table 4**Error! Reference source not found.** shows that a sizeable share of errors in dollars received (3.0-6.8 percent) is due to imputations to non-recipients. Between 25 and 75 percent of imputed dollars are assigned to households that do not receive the program. At the same time, a similar share (2.9-7.2 percent) is missed by the imputations as the second row of Table 4**Error! Reference source not found.** shows, which confirms that the low net imputation error masks substantial imputation error at the household level. On the positive side, net error due to mis-imputing receipt status is low both in absolute terms and as a fraction of total item non-response error. The fourth row of Table 4**Error! Reference source not found.** shows that item non-response error is mainly due to systematic errors in the amounts imputed when receipt status is correctly imputed. These underimputed amounts are particularly relevant for PA in the SIPP, where they make up 67 percent of the sizeable total error. PA amounts in the CPS are an exception, as they are imputed correctly on average.

Decomposing measurement error into error in a binary receipt variable and a continuous amount received variable is important, because surveys often collect this information from two separate questions. For example, surveys commonly ask whether respondents have income from employment or a transfer program and only ask for the amount received if the answer is positive. Consequently, decomposing these sources of errors also sheds light on different aspects of survey design. Table 4**Error! Reference source not found.** shows that, in contrast to imputation, misreporting of receipt status among respondents accounts for a larger share of net underreporting than misreporting of amounts. This error is most pronounced for the CPS, where 92 percent of SNAP underreporting and almost three quarters of the error for PA are due to errors in receipt status. For the ACS and SIPP, the share of measurement error due to misreporting of receipt status is around 60 percent. The error in dollars missed due to errors in receipt status is driven by true recipients failing to report receipt, which confirms previous evidence that the number of recipients is underestimated (Marquis and Moore 1990; Taeuber et al. 2004; Lynch et al. 2007; Meyer, Mittag, and Goerge 2018; Celhay, Meyer and Mittag 2017b). Yet, we also find a sizeable offsetting effect due to dollars reported by non-recipients. This effect of false positives is low at 2 percent of average dollars paid for SNAP in the CPS, but is more pronounced for PA in ACS, where false positives amount to up to 14 percent of the survey target. In the other surveys, non-recipients report around 6 percent of the survey target. Consequently, aggregate reporting rates provide a good measure of error for SNAP, but in our surveys they are insufficient for PA.

An important caveat when examining overreporting is that we correct for payments of NY SNAP and PA to individuals residing in other states according to the surveys, but cannot correct for payments from other states received by households in our sample. If these households report receipt, some "false positives" are in fact true recipients. The subtractions for out of state payments in Table 1 are smaller than dollars reported by false positives (except for SNAP in the CPS), but may account for a sizeable fraction of receipt by those we classify as overreporters here. Nevertheless, this comparison is at best suggestive. Even though net migration into NY among the poor is close to zero, it would also require receipt among migrants into and out of NY to be symmetric and heavily depends on their reporting rate.[24]

Finally, amounts received by those who correctly report receipt status are also underreported on average in all cases. For SNAP, net underreporting of amounts is relatively small at 2.7 percent in the CPS and 4.3 percent in the SIPP. For PA, underreporting of amounts is sizeable. It is surprisingly similar across surveys at around 20 percent of the target amount. A potential reason why there is net underreporting of amounts for PA, but not SNAP, is that in many cases, a fraction of the PA payment is made directly to third parties such as landlords and utility companies. Respondents may fail to include this part of the payment in their reported amount received.[25]

5. Extensions

Our approach of using data combination to estimate and decompose TSE can be extended in many ways. We focus on the error in average dollars paid, but our data also allow us to do the same decomposition for the number of program recipients or assistance units. Data on other government programs could be used to estimate and decompose TSE in receipt and amounts received from, among others, social security, Supplemental Security Income or unemployment benefits. More generally, our approach of estimating TSE and its components can be applied whenever an accurate record corresponding to a survey variable can be obtained for the entire population of interest and linked. Such measures could come from administrative data on variables such as education or health insurance status. Data sources that cover the entire population or a large share of it are increasingly becoming available from administrative records

[24] These payments from out of state are not included in our survey target, since they are not included in the NY administrative data. Technically, we study to what extent the NY survey sample captures dollars paid by NY OTDA, so that we should exclude amounts paid by other states that are reported by households in NY, which we cannot do. We thereby understate how much the surveys fall short of their target (TSE is negative). It does not affect our estimate of generalized coverage error. Our estimates of item non-response and measurement error are slightly too high, so the problem of underreporting is more severe than our numbers suggest.

[25] Our communications with current and former New York OTDA employees confirmed that such payments to third parties are common, but we were not able to learn the share of cases where they occur.

such as tax filings. The approach can also be applied to business surveys, since government records often contain data on subsidies paid to firms, their number of employees, or the capital that they raised. It may also be possible to extend our approach to other variables using records from private companies, such as credit card records or utility records to analyze consumption, or medical records to analyze health.

One could also extend our approach to study survey error for subpopulations, such as single parents or the elderly. Recent validation studies provide evidence that the accuracy of imputations and reporting varies with demographic characteristics (Bollinger and David 2001; Celhay, Meyer and Mittag 2017b; Meyer and Mittag, forthcoming). The decomposition we propose could provide further detail to gauge the accuracy of survey-based studies of subpopulations. The question to what extent surveys cover populations such as single parents (Meyer and Sullivan, 2008) or extremely poor households (Edin and Schaefer, 2015) is crucial not only for scientific research, but also for anti-poverty policies and the targeting of transfer programs. Our method of analyzing coverage by means of data linkage could be extended to answer these questions. As we mention in section 2.b, record linkage provides us with a powerful tool to analyze survey representativeness. The general principle of linking records that cover the (sub-)population of interest and then comparing statistics from these records to (weighted) statistics from the linked cases to examine coverage is widely applicable. This approach requires data on the entire population of interest and reliable linkage, but neither relies on having an accurate nor a comparable measure of a survey variable in the population records. If the population data only provide a noisy proxy, we can still analyze unit non-response by examining whether the distribution of the proxy depends on response status. See Bee, Gathright and Meyer (2017) for an example and further discussion. . In the extreme case where the proxy is completely uninformative, e.g. because the population data only contain a list of units without further information, we can still estimate the population size of linked units and compare it to the actual population size in the population data. We also do not need to be able to identify the population of interest in the survey. The comparison is valid as long as the population data can be restricted or adjusted to match the population the survey intends to cover (or, more generally, the subpopulation of interest, such as program recipients).

As our more detailed decomposition in equation (14) shows, one can also extend the analysis with additional data. For generalized coverage error, for example, we combine sources of error such as frame error, unit non-response and re-weighting error. The contributions of these sources of error to TSE can be separated if one is able to link administrative data directly to the sampling frame, rather than to the final survey data as in Bee, Gathright and Meyer (2017). Frame error is the difference between total X linked to the survey frame and the total in the population data. Unit non-response error can be estimated from the records linked to unit non-respondents in the survey frame and the effect of the weight adjustment is the difference in the estimate using the adjusted and the base weights. Further decomposing error that arises later in the survey process is simpler to implement. After linking the data, error components due to survey post-processing can be estimated if the original and the edited variable are available.

Finally, the error decomposition could be extended to other parameters. We focus on the average error in dollars paid, but the same methods could be applied to parameters such as the variance of amounts or the mean squared error of estimates. For example, TSE in the variance of the transfer amount per household could be estimated as above. Extending the decomposition to the variance may be of interest for substantive reasons, such as estimating the variance of income to study inequality. In our case, it would allow us to examine to what extent survey estimates reflect the contribution of government transfers to the reduction in inequality.

6. Conclusions

We argue that to understand the growing problem of survey error and to ultimately reduce it, we need a framework to measure the extent, sources, and likely impact of non-sampling error. In survey methodology, the Total Survey Error Framework plays a key role in conceptualizing the size and impact of different sources of error at the design stage. The same framework could be used to measure, analyze, and reduce survey error during or after data collection. Yet, TSE and its components are not routinely estimated, because it is usually infeasible or prohibitively costly. We show that data combination can provide an inexpensive way to routinely implement an empirical TSE decomposition. We define measures of TSE and its key components that can be estimated using data combination. We demonstrate the feasibility and usefulness of this framework by estimating and decomposing TSE in average dollars paid by two important government transfer programs in three major U.S. household surveys that we link to administrative records.

We find that TSE is substantial and leads to understatement of transfer dollars received in all three surveys, with larger overall error for PA than for SNAP. Its size and composition varies across surveys and its composition varies across similar questions in the same survey. Measurement error is by far the largest source of error and mainly due to underreporting of receipt. Our estimates of generalized coverage error and item non-response error are much smaller and often bias estimates in opposing directions. The TSE due to these two components combined is less than 10 percent of the survey target in all cases. From a methodological perspective, these results demonstrate that data combination can provide a powerful tool to empirically implement the TSE framework and to jointly measure multiple sources of non-sampling error. Databases covering the entire (survey) population are increasingly available and linking them to survey data has become cheaper and more accurate, so that the measures of error we propose could be produced on a regular basis. These methods would help survey producers improve, and survey users better understand, survey accuracy.

For survey producers, understanding the importance of each error source is crucial to reduce overall error in a cost-effective manner. Methods to reduce most error sources are known, but costly, so it is important to understand the importance of each error source to reduce survey error in a cost-effective manner. See e.g. Spencer (1985) for a cost-benefit framework of optimal data quality. For example, the much larger extent of measurement error suggests that the marginal cost of reducing this error source may be lower than further reducing non-response error. The decomposition can also provide survey producers with insights into the nature and possible causes of error and hence point to potential remedies. For example, our approach to analyze (generalized) coverage error provides a simple and powerful tool to study a source of error that is difficult to analyze. The results can provide survey producers with information on the populations that the survey fails to capture and hence how to improve survey coverage. Finally, routine implementation of a framework to measure and decompose TSE would allow survey producers to monitor changes in the extent of errors and thereby evaluate design choices. Tools to reduce survey error are available, but the variation in error within and across surveys suggests that their effectiveness in a specific case is hard to assess ex ante. Continuously monitoring survey error would allow survey agencies to evaluate their design choices and better tailor measures to reduce error and be less expensive.

For survey users, measures of TSE and its components are important to understand which surveys and which questions are reliable and thereby help them to improve the accuracy of their estimates. For example, knowing which populations are likely (not) covered by the survey is important to gauge which populations the survey can reliably measure. Our finding of high coverage rates of program recipients is encouraging for studies of poor households, although further research is required to examine whether our findings extend to poor non-recipients. Estimates of item non-response error can help survey users decide whether they should use the survey imputations. Our results suggest that imputed values can reduce the consequences of item non-response when estimating population means. This result is encouraging in light of the evidence that imputations frequently fail to capture key correlations (Hirsch and Schumacher 2004; Bollinger and Hirsch 2006) and induce substantial error at the individual level (Celhay, Meyer and Mittag 2017b; Meyer, Mittag, and Goerge 2018). At the same time, our error decomposition points to substantial error at the household level, which makes using the imputed observations in multivariate analyses questionable. In a similar vein, prior studies have pointed out that despite the high error rates, multivariate models of the determinants of program receipt still tend to get the signs of key coefficients right. The large and negative bias from measurement error that we find emphasizes that this robustness is unlikely to extend to estimated receipt rates.

Extending the decomposition as discussed in section 5 may yield further benefits. Our approach of using data combination to estimate average (non-sampling) error in survey means could be extended to other statistics such as mean squared error and variance. With further data on additional error sources or other survey variables, it could be applied to a wider range of error components and other survey variables known to suffer from error. Thereby, this study underlines the value of data combination and linkage as a tool to analyze survey quality and shows the potential of this approach to mitigate the critical problem of survey data quality.

References

Alho, Juha M. and Bruce D. Spencer 1985. “Uncertain Population Forecasting.” *Journal of the American Statistical Association* 80, 306-314.

Alwin, Duane F. 2007. *Margins of Error: A Study of Reliability in Survey Measurement*. New York: Wiley.

Alwin, Duane F., 1991. Research on survey quality. *Sociological Methods & Research*, *20*(1): 3-29.

Bee, C. Adam, and Joshua Mitchell. 2017. “The Hidden Resources of Women Working Longer: Evidence from Linked Survey-Administrative Data.” In *Women Working Longer: Increased Employment at Older Ages*, eds. Claudia Goldin and Lawrence F. Katz. Chicago: University of Chicago Press.

Bee, C. Adam, Graton Gathright and Bruce D. Meyer. 2017. “Bias from unit non-response in the measurement of income in household surveys.”

Biemer, Paul P., and Lars Lyberg. 2003. *Introduction to Survey Quality.* New York: Wiley.

Biemer, Paul P. 2009. ‘‘Measurement Error in Sample Surveys.’’ In *Handbook of Statistics 29A*, eds. Danny Pfefferman and C.R. Rao, Chapter 12, 281--315. Amsterdam: North-Holland.

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

Bollinger, Christopher R., and Martin H. David. 2001. “Estimation with response error and nonresponse: food-stamp participation in the SIPP.” *Journal of Business & Economic Statistics* 19(2): 129-141.

Bollinger, Christopher R., and Barry T. Hirsch. 2006. “Match Bias Due to Earnings Imputation: The Case of Imperfect Matching.” *Journal of Labor Economics* 24 (3).

Bollinger, Christopher R., Barry T. Hirsch , Charles M. Hokayem, and James P. Ziliak. forthcoming. “Trouble in the Tails? What We Know about Earnings Nonresponse Thirty Years after Lillard, Smith, and Welch.” *Journal of Political Economy.*

Bound, John, Charles Brown, and Nancy Mathiowetz. 2001. “Measurement error in survey data.” In *Handbook of Econometrics*. Vol. 5, eds. James J. Heckman and Edward Leamer, Chapter 59, 3705 – 3843. Amsterdam: Elsevier.

Brackstone, Gordon. 1999. ‘‘Managing Data Quality in a Statistical Agency.’’ *Survey Methodology* 25(2):139–49.

Celhay, Pablo, Bruce D. Meyer and Nikolas Mittag. 2017a. “What Leads to Measurement Error? Evidence from Reports of Program Participation in Three Surveys.” Unpublished Manuscript.

Celhay, Pablo, Bruce D. Meyer and Nikolas Mittag. 2017b. “Errors in Reporting and Imputation of Government Benefits and Their Implications.” Unpublished Manuscript.

de Leeuw, Edith, and Wim de Heer. 2002. "Trends in Household Survey Nonresponse: A Longitudinal and International Comparison." In *Survey Nonresponse*, ed. Robert M. Groves, Don A.Dillman, John L. Eltinge, and Roderick J. A. Little, pp. 41-54. New York: Wiley.

Edin, Kathryn J., & Shaefer, H. Luke. 2015. *$2.00 a day: Living on almost nothing in America*. Boston: Houghton Mifflin Harcourt.

Eurostat. 2000. ‘‘Assessment of the Quality in Statistics.’’ Eurostat General/Standard Report, Luxembourg, April 4–5.

Gathright, Graton M. R., and Crabb, Taylor A. 2014. Reporting of SSA Program Participation in SIPP. Working Paper, U.S. Census Bureau.

Groves, Robert M. 2004. *Survey errors and survey costs*. Vol. 536. New York: John Wiley & Sons.

Groves, Robert M., and Magilavy, Lou J. (1986). Measuring and explaining interviewer effects in centralized telephone surveys. *Public Opinion Quarterly*, *50*(2), 251-266.

Groves, Robert, Floyd Fowler, Mick Couper, Eleanor Singer, and Roger Tourangeau. 2004. *Survey Methodology*. New York: Wiley.

Groves, Robert M., and Lars Lyberg. 2010. “Total Survey Error: Past, Present, and Future.” *Public Opinion Quarterly*, 74(5): 849–79.

Hirsch, Barry T., and Edward Schumacher. 2004. “Match Bias in Wage Gap Estimates Due to Earnings Imputation.” *Journal of Labor Economics*, 22(3): 689–722.

Hokayem, Charles, Christopher R. Bollinger, and James P. Ziliak. 2015. “The Role of CPS Nonresponse in the Measurement of Poverty.” *Journal of the American Statistical Association*, 110(511), 935-945.

Kirlin, John A., and Wiseman, Michael 2014. “Getting it Right, or at Least Better: Improving Identification of Food Stamp Participants in the National Health and Nutrition Examination Survey.” Working Paper.

Lynch, Victoria, Dean M. Resnick, Jane Stavely, and Cynthia M. Taeuber. 2007. “Differences in Estimates of Public Assistance Recipiency Between Surveys and Administrative Records.” U.S. Census Bureau, Washington, D.C.

Manski, Charles. 2015. “Communicating Uncertainty in Official Economic Statistics: An Appraisal Fifty Years after Morgenstern.” *Journal of Economic Literature,* 53(3), 631-653.

Marquis, Kent H, and Jeffrey C Moore. 1990. “Measurement Errors in SIPP Program Reports.” U.S. Census Bureau.

Meyer, Bruce D., Nikolas Mittag and Robert Goerge. 2018. “Errors in Survey Reporting and Imputation and Their Effects on Estimates of Food Stamp Program Participation.” NBER Working Paper No. 25143.

Meyer, Bruce D., and Nikolas Mittag. forthcoming. “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*.

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

Meyer, Bruce D. and James X. Sullivan. 2008. “Changes in the Consumption, Income, and Well-Being of Single Mother Headed Families.” *The American Economic Review*, 98(5), 2221-2241.

Mulry, Mary H. and Bruce D. Spencer. 1988. “Total Error in the Dual System Estimator: The 1986 Census of Central Los Angeles County.” *Survey Methodology* 14, 241-263.

Mulry, Mary H. and Bruce D. Spencer. 1991. “Total Error in PES Estimates of Population: The Dress Rehearsal Census of 1988.” *Journal of the American Statistical Association* 86, 839-854.

Mulry, Mary H. and Bruce D. Spencer. 1993. “Accuracy of the 1990 Census and Undercount Adjustments.” *Journal of the American Statistical Association* 88, 1080-1091.

Mulry, Mary H. 2007 "Summary of accuracy and coverage evaluation for the US Census 2000." *Journal of Official Statistics* 23, 345.

Mulry, Mary H., and B. D. Spencer. 2001. “Accuracy and coverage evaluation: Overview of total error modeling and loss function analysis.” *DSSD Census 2000 Procedures and Operations Memorandum, Series B-19*. Washington, D.C.: U.S. Census Bureau.

Nicholas, Joyce and Michael Wiseman. 2009. “Elderly Poverty and Supplemental Security Income.” *Social Security Bulletin*, 69: 1, 45-73.

NORC. 2011. “Assessment of the US Census Bureau’s Person Identification Validation System.” NORC at the University of Chicago Final Report presented to the US Census Bureau.

OECD. 2003. “Quality Framework and Guidelines for Statistical Activities.” Version 2003/1.

Ridder, Geert, and Moffitt, Robert 2007. “The econometrics of data combination.” In *Handbook of Econometrics*. Vol. 6B, eds. James J. Heckman and Edward Leamer, Chapter 75, p. 5469-5547. Amsterdam: Elsevier.

Spencer, Bruce D. 1985. “Optimal Data Quality.” *Journal of the American Statistical Association* 80, 564-573.

Taeuber, Cynthia, Dean M. Resnick, Susan P. Love, Jane Stavely, Parke Wilde, and Richard Larson. 2004. “Differences in Estimates of Food Stamp Program Participation Between Surveys and Administrative Records” Working Paper, U.S. Census Bureau.

U.S. Census Bureau. 2006. “Design and Methodology: Current Population Survey.” U.S. Census Bureau.

U.S. Census Bureau. 2008. “Survey of Income and Program Participation: User’s Guide.” U.S. Census Bureau.

U.S. Census Bureau. 2012. “Changes to ACS Group Quarters Small Area Estimation.” ACS User Note retrieved from https://www.census.gov/programs-surveys/acs/technical-documentation/user-notes/2011-01.html

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

Wagner, Deborah, and Mary Layne. 2014. “The Person Identification Validation System (PVS): Applying the Center for Administrative Records Research and Applications’ (CARRA) Record Linkage Software.” U.S. Census Bureau.

Wooldridge, Jeffrey M. 2007. “Inverse Probability Weighted Estimation for General Missing Data Problems.” *Journal of Econometrics*, 141(2): 1281–1301.

Appendix 1 – Survey Data

Survey population and our sample of interest

Our population of interest is the entire population residing in households, which is defined (following the U.S. Census Bureau) by living in a residential structure that is not a group quarter. Those not in residential structures are largely the street homeless. Group quarters definitions are not entirely comparable across surveys, so we exclude those residing in group quarters and focus on the population in households. Specifically, the ACS includes a household and a group quarters sample. We only use the former, which covers our exact population of interest. The CPS and SIPP include the population residing in households as well as a small sample from non-institutional civilian group quarters, which we exclude from our sample.

The surveys are often used to obtain estimates for the civilian non-institutional population (persons 16 years of age and older residing in the 50 states and DC who do not live in institutions and are not on active duty in the Armed Forces). This population of individuals slightly differs from the individuals living in the household population we examine, which additionally includes those younger than 16 and military members living in households, but excludes individuals living in non-institutional group quarters.

Adjustments to Survey Sampling Weights

All three surveys take unit non-response into account by adjusting base weights to make demographic characteristics of the sample match population statistics. All three surveys are household samples, so the base weights are the inverse of the probability of a household being sampled. In a first step, all three surveys create individual level weights by assigning these households weights to all individuals in the household. The adjusted household weights are then obtained from these individual weights after a series of adjustments that differ between surveys. The ACS adjusts individual base weights in two stages (U.S. Census Bureau, 2014, p. 142-161). The first stage adjusts for unit non-response, the second stage makes weighted estimates of the number of households and persons by age, sex, race and Hispanic origin match control totals. Both adjustments are made by county or group of less populated counties. Our data include the weights after the first adjustment and the final weights, but not the base weights. The CPS ASEC uses a complex multi-stage procedure to adjust the individual base weights. It includes adjustments for changes to the interview list after sampling, combining households from multiple waves (rotation groups) of the monthly basic CPS and the inclusion of the State Children's Health Insurance Program sample in the ASEC. See U.S. Census Bureau (2006) chapters 10 and 11 for a detailed description and page 11-9 for a diagram. Similarly, the SIPP not only adjusts the individual base weights for non-response, but also for changes to the interview list after sampling and mobility during the panel. It then applies a post-stratification adjustment to correct for departures from known population totals (U.S. Census Bureau, 2006, page 8-6 to 8-11). All three surveys then obtain the household weights we use from the adjusted individual weight of the householder. The ACS adjusts these household weights to make the sum of the weights, P^S, match the number of households by state and demographic characteristics (U.S. Census Bureau 2014, parts 11.7-11.10). The CPS and the SIPP use the individual weights of the householder without controlling the size of the household population. The CPS makes a small adjustment to adjust the gender ratio of married couples. The number of households in NY implied by the SIPP weights is very close to the corresponding numbers from the ACS and hence official estimates, but the CPS weights suggest a household population that is about 5 percent larger.

Imputation Methods

All three surveys impute missing values using a hot deck procedure. In short, the hot deck sorts observations into cells based on categorical variables reported in the survey. If a variable is missing for an observation, the value of a respondent from the same cell is assigned to this observation instead. The details of the implementation vary across surveys and variables. For example, for SNAP, the ACS constructs cells based on few demographic characteristics (family type, presence of children, poverty status, and the race of the reference person), but incorporates detailed geographic information by only using values from the same state and assigning the value of the most recent respondent in the corresponding cell at the smallest geographic level available. In contrast, the CPS hot deck for SNAP classifies households into a much larger number of cells (648), but at the national level. Contrary to SNAP, the CPS imputes PA jointly with other missing income components from a single donor. The SIPP hot deck uses a comparable number of cells (864) to impute SNAP at the national level, but also incorporates some geographic information and restricts imputed values to come from the same wave. For more detailed descriptions see US Census Bureau (2006, 2008 and 2014) for the CPS, SIPP, and ACS respectively and Meyer, Mittag and Goerge (2018) for a summary.

Appendix 2 – Adjusting for Unlinkable Survey Households Using Inverse Probability Weighting

We link both the survey data and the administrative data to a third data source. Thus, we can distinguish between cases in which no match was found because the individual is not present in the other data source and cases in which no match was found because the PVS could not find a PIK. We can adjust for such unlinkable records in the administrative data by subtracting the total amount paid to records without a PIK from the survey target. Missing PIKs in the survey data introduce an additional source of linkage error, because our linked data excludes survey households for which no household member has a PIK. Contrary to missing PIKs in the administrative data, we do not know x_i^A, the amount paid according to the administrative records, for these households. Instead, we have the detailed demographic information of the surveys, so we adjust for this source of linkage error by inverse probability weighting (IPW).

IPW and the assumptions it requires are discussed, among others, in Wooldridge (2007). In short, IPW adjusts for missing data by estimating the probability of an observation being missing (conditional on covariates Z) and multiplies the survey weights by the inverse of this probability. That is, in our case the estimated weight adjustment for missing households due to unlinkable survey households is $\widehat{w}_i^{IPW} = \frac{1}{\widehat{\Pr}(PIK=1|Z)}$ as defined above. As Wooldridge discusses, the adjustment is consistent under an assumption similar to MAR (missing conditionally at random). In particular, our application is a simple case covered by theorem 3.1 of Wooldridge (2007), so that the key assumptions for consistency are assumption 3.1 and 3.2 in Wooldridge (2007). For the general case, assumption 3.1 requires that the probability of being included in the data (i.e. having a PIK) does not depend on the variables in the model of interest, W, conditional on the variables used to estimate the weights, Z, i.e. $\Pr(PIK = 1|Z, W) = \Pr(PIK = 1|Z)$. We are only interested in estimating the mean of a variable, so we only need the weaker assumption that the mean of our outcome of interest conditional on Z does not depend on PIK-status: $\mathbb{E}(W|PIK = 1, Z) = \mathbb{E}(W|Z)$. Assumption 3.2 requires us to be able to estimate $\Pr(PIK = 1|Z)$ by maximum likelihood.

To estimate the weight adjustment, we use Probit models with PIK, the indicator whether someone in the household has a PIK, as the dependent variable. That is, we assume $\Pr(PIK = 1|Z) = \Phi(Z\beta)$ and estimate β by standard maximum likelihood. The covariates Z that we include in these models differ slightly between surveys, but always include household type, the number of adults and children present, the number of employed members, household income relative to the poverty line and whether the household is in a rural area. We also include information on the household head, specifically age, education and race/ethnicity categories and indicators for gender and disability of the household head. In the ACS, we additionally include dummies whether English is the only language spoken in the household and whether the household head speaks English poorly or is not a U.S. citizen. We estimate separate Probit models for each survey in each year. For our main analyses, we use weights obtained from estimating the Probit models for our samples of interest only, i.e. for households residing in NY. We use separate annual Probit models that include the entire U.S. survey sample (and state fixed effects) to construct the weights in our adjustment for payments to households outside of NY. To adjust the weights for individuals in the linked ACS group quarters sample, we estimate separate annual Probit models using the NY sample at the individual level. To do so, we define individual-level counterparts of the covariates Z when necessary.

Table 1 - Total SNAP and PA Payments that Should be Covered by the Linked NY Household Sample in Three Surveys

	SNAP						**Public Assistance**					
Survey	ACS		CPS		SIPP		ACS		CPS		SIPP	
Time period covered*	2008-2012		2007-2012		2007-2012		2008-2012		2008-2013		2007-2012	
	$M	%	$M	%	$M	%	$M	%	$M	%	$M	%
Payments in administrative microdata	4,345		4,264		4,444		1,717		1,711		1,724	
Subtraction of payments not covered by household samples												
Payments to non-residential/homeless population	-52	-1.2%	-51	-1.2%	-48	-1.1%	-8	-0.5%	-8	-0.5%	-7	-0.4%
Estimated payments to group quarters	-233	-5.4%	-227	-5.3%	-210	-4.7%	-166	-9.7%	-165	-9.7%	-148	-8.6%
Subtraction of payments not covered by linked NY data												
Unlinkable administrative records (no PIK)	-21	-0.5%	-20	-0.5%	-18	-0.4%	-34	-2.0%	-30	-1.8%	-28	-1.6%
Estimated payments linked to residents of other states	-135	-3.1%	-141	-3.3%	-97	-2.2%	-43	-2.5%	-94	-5.5%	-21	-1.2%
Total subtraction	-441	-10.1%	-438	-10.3%	-373	-8.4%	-251	-14.6%	-298	-17.4%	-204	-11.9%
Total payments covered by linked NY data	3,904	89.9%	3,826	89.7%	4,031	90.7%	1,466	85.4%	1,413	82.6%	1,497	86.9%

Notes: We use the survey sample of households by excluding group quarters. We calculate payments to the non-residential population and to unlinkable administrative records from the administrative microdata. Estimated payments to group quarters are based on the 2008-2010 ACS NY group quarters sample and extrapolated for 2011-2012. Estimated payments linked to residents of other states are estimated from NY OTDA records linked to the non-NY sample of each survey. All dollar amounts are average annual amounts in millions of dollars. All percentages are calculated as a percent of total dollars paid according to the administrative microdata. We use survey weights adjusted for PIK probability for the estimates from the linked data for payments to group quarters and residents of other states.

* A survey year in the ACS combines two calendar years, so the 2008 data partly refer to 2007, see footnote 14. The CPS period refers to calendar years. The time period covered by the SIPP combines parts of the 2004 SIPP panel (calendar year 2007) and the 2008 SIPP panel (August 2008 - December 2012).

Table 2 - TSE in Average Annual SNAP and PA Dollars Paid to NY Households for Three Surveys

	SNAP			Public Assistance		
Survey	ACS	CPS	SIPP	ACS	CPS	SIPP
Average dollars paid per NY household (survey target)	543	532	522	204	197	194
Estimated average dollars reported by NY households		372	500	88	79	80
Total survey error in dollars per household		-160	-23	-116	-118	-114
Total survey error in percent of survey target		-30.0%	-4.3%	-57.1%	-59.8%	-58.9%

Notes: The first row contains average payments covered by the linked household data from Table 1 divided by the number of households in the population. See footnote 23. We use the linked data and survey weights adjusted for PIK probability to estimate average dollars reported in the second row. Average dollars are annual dollars per household obtained by dividing the sum of annual payments by the sum of annual number of households for the survey periods in the ACS and CPS. In the SIPP, we use the sum of monthly payments divided by the average monthly number of households. The third row is the difference between row one and two, the fourth row is the ratio of the third and first row. The ACS does not ask for SNAP amounts.

Table 3 - Total Survey Error and its Components as Share of Average Dollars Paid per Household

	SNAP			Public Assistance		
	ACS	CPS	SIPP	ACS	CPS	SIPP
Generalized Coverage Error	-1.0%	5.2%	10.7%	-4.5%	6.4%	10.0%
Item non-response error		-2.7%	-1.1%	-2.1%	2.3%	-3.1%
Measurement error		-32.5%	-13.9%	-50.5%	-68.5%	-65.8%
Total net survey error		-30.0%	-4.3%	-57.1%	-59.8%	-58.9%

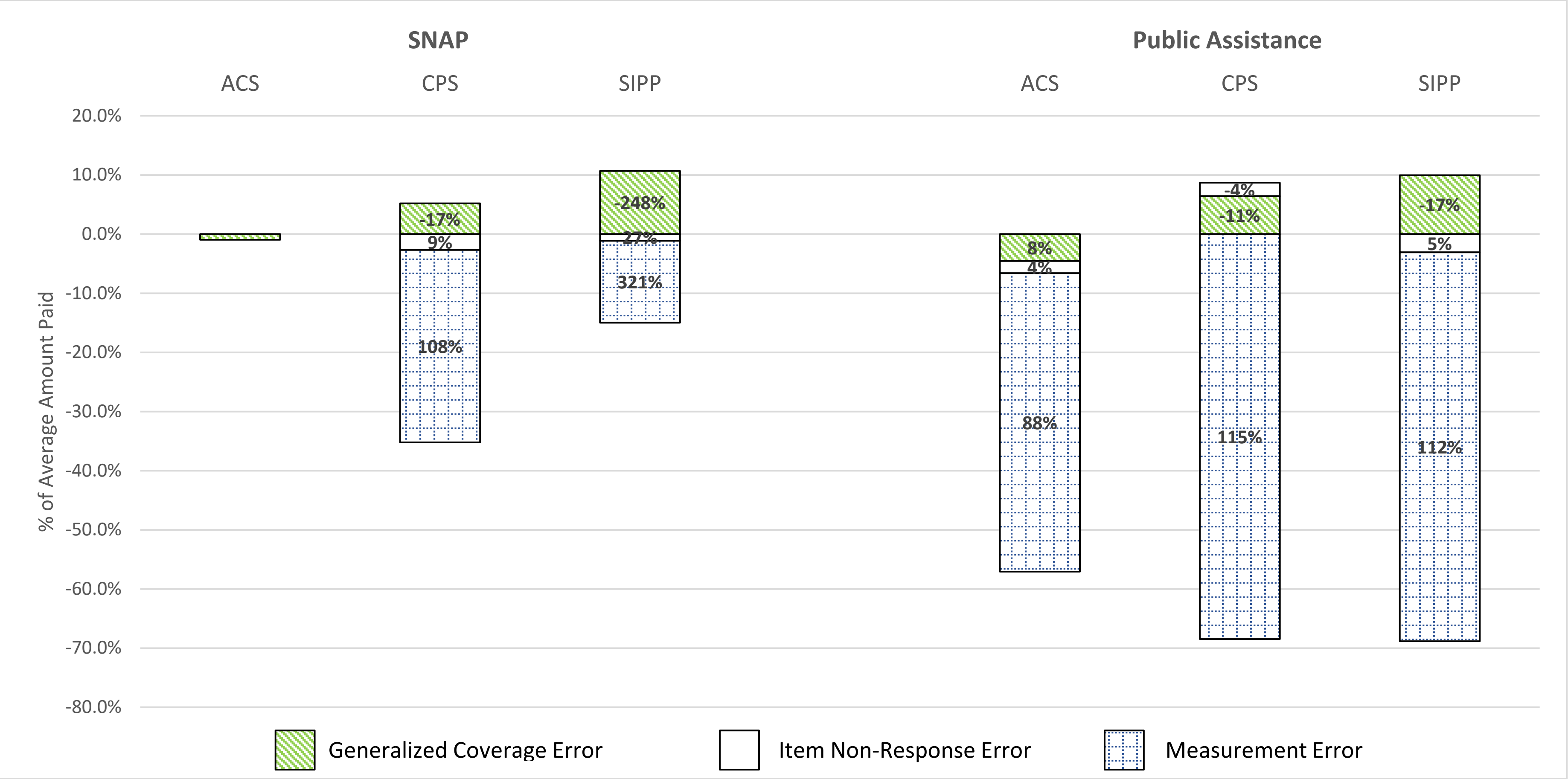
Notes: All amounts are in percent of average dollars paid to households covered by the linked NY data. All estimates are based on the linked NY sample using survey weights adjusted for PIK probability. Generalized coverage error is estimated average dollars paid in the linked administrative data minus average dollars paid in the unlinked administrative data. Item non-response and measurement error are the differences between estimated average dollars paid according to the survey variable and the administrative variable in the linked data divided by the size of the overall population for item non-respondents and respondents respectively. Total net survey error is the sum of the three components. The ACS does not ask for SNAP amounts, so we can only estimate coverage error. The SIPP is not claimed to be representative of NY state, so the estimates of coverage error for the SIPP should be interpreted with caution. See notes to Table 1 and Appendix 1 for further information on the samples and time periods.

Table 4 - Sources of Item Non-Response Error and Measurement Error as Shares of Average Dollars Paid

	SNAP		Public Assistance		
	CPS	SIPP	ACS	CPS	SIPP
	Item Non-Response Error				
True non-recipients imputed as recipients (false positives)	6.3%	3.0%	5.8%	6.8%	4.1%
True recipients imputed as non-recipients (false negatives)	-7.2%	-2.9%	-4.9%	-4.5%	-5.1%
Net error due to mis-imputed receipt status	*-0.9%*	*0.1%*	*0.9%*	*2.3%*	*-1.0%*
Net error in imputed amounts when receipt is correctly imputed	-1.8%	-1.2%	-3.0%	0.0%	-2.1%
Net item non-response error	*-2.7%*	*-1.1%*	*-2.1%*	*2.3%*	*-3.1%*
	Measurement Error				
True non-recipients reporting receipt (false positives)	1.9%	6.4%	14.0%	6.1%	5.9%
True recipients not reporting receipt (false negatives)	-31.8%	-16.0%	-42.8%	-56.4%	-47.5%
Net error due to misreported receipt status	*-29.8%*	*-9.6%*	*-28.8%*	*-50.3%*	*-41.5%*
Net error in amounts by reporting true recipients	-2.7%	-4.3%	-21.7%	-18.2%	-24.2%
Net measurement error	*-32.5%*	*-13.9%*	*-50.5%*	*-68.5%*	*-65.8%*

Notes: All amounts are in percent of average dollars paid to households covered by the linked NY data. All estimates are based on the linked NY sample using survey weights adjusted for PIK probability. The first row of each panel contains the survey estimates of dollars paid to those receiving the program according to the survey variable, but not the linked administrative variable. The second row contains dollars paid according to administrative records to those receiving the program according to the administrative variable, but not the survey variable. The fourth row is the difference between the average amount paid according to the survey and the administrative measure among those receiving the program according to both the administrative and the survey variable. The third and fifth rows are the sum of the respective two rows above. See notes to table 1 and Appendix 1 for further information on the samples and time periods.

Figure 1: TSE Components in Percent of Average Dollars Paid and as Shares of TSE



Note: The size of the bars indicates the error amount as a percentage of average dollars paid to NY households in the time period covered by the survey. The numbers in the bars are the error components as a percentage of total survey error for the respective survey and program. See Table 1 for definitions and further notes.