WORKING PAPER · NO. 2022-11

What Leads to Measurement Errors? Evidence from Reports of Program Participation in Three Surveys

Pablo A. Celhay, Bruce D. Meyer, and Nikolas Mittag

JANUARY 2022

Becker Friedman Institute
FOR ECONOMICS AT UCHICAGO

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

WHAT LEADS TO MEASUREMENT ERRORS? EVIDENCE FROM REPORTS OF PROGRAM PARTICIPATION IN THREE SURVEYS

Pablo A. Celhay
Bruce D. Meyer
Nikolas Mittag

January 2022

This paper, which has been subject to a limited Census Bureau review, is released to inform interested parties of research and to encourage discussion. Any opinions and conclusions expressed herein are those of the authors and do not represent the views of the U.S. Census Bureau or the New York Office of Temporary and Disability Assistance (OTDA). The Census Bureau has reviewed this data product for unauthorized disclosure of confidential information and has approved the disclosure avoidance practices applied to this release, authorization dates: October 26, 2016 and October 6, 2017. We are grateful for the assistance of current and former Census Bureau employees including David Johnson, Amy O'Hara, Graton Gathright, Frank Limehouse, Trent Alexander, and Joey Morales and New York OTDA employees Dave Dlugolecki and George Falco. The authors also thank Dan Black, Charles Brown, Jeff Grogger, Derek Wu, and participants in talks at The University of Chicago, Universidad Adolfo Ibañez, IAAE, EEA-ESEM, AEA, CFE and JSM for helpful comments. We appreciate the financial support of the Alfred P. Sloan Foundation, the Russell Sage Foundation, the Charles Koch Foundation, the Menard Family Foundation, and the American Enterprise Institute. Mittag is also grateful for financial support from the Czech Science Foundation (through grants no. 16-07603Y and 20-27317S) and the Czech Academy of Sciences (through institutional support RVO 67985998).

What Leads to Measurement Errors? Evidence from Reports of Program Participation in Three Surveys

Pablo A. Celhay, Bruce D. Meyer, and Nikolas Mittag

January 2022
JEL No. C81,D31,I32,I38

ABSTRACT

Measurement errors are often a large source of bias in survey data. Lack of knowledge of the determinants of such errors makes it difficult for data producers to reduce the extent of errors and for data users to assess the validity of analyses using the data. We study the determinants of reporting error using high quality administrative data on government transfers linked to three major U.S. surveys. Our results support several theories of misreporting: Errors are related to event recall, forward and backward telescoping, salience of receipt, the stigma of reporting participation in welfare programs and respondent's degree of cooperation with the survey overall. We provide evidence on how survey design choices affect reporting errors. Our findings help survey users to gauge the reliability of their data and to devise estimation strategies that can correct for systematic errors, such as instrumental variable approaches. Understanding survey errors allows survey producers to reduce them by improving survey design. Our results indicate that survey producers should take into account that higher response rates as well as collecting more detailed information may have negative effects on survey accuracy.

Pablo A. Celhay
School of Government
and Economics Institute
Pontificia Universidad Católica de Chile
Avda. Vicuña Mackenna 4860
Macul, Santiago
Chile
pablocelhay@gmail.com

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
CERGE-EI
Politických vězňů 7
Praha
Czech Republic
nikolas.mittag@cerge-ei.cz

I. Introduction

Surveys are one of the most important tools for empirical work in economics and other social sciences to examine human behavior. Economic research, which often emphasizes analyzing what people do rather than what they say, usually analyzes what people report that they do. Government operations rely on household surveys as a main source of data used to produce official statistics, including unemployment, poverty, and health insurance coverage rates. Academic researchers also rely on these large-scale surveys heavily or conduct their own surveys, for example when running randomized control trials. This dependence of science and governance on reported information raises the question whether and when survey reports accurately reflect truth. Unfortunately, survey data have been found to contain error in a wide range of settings. For example, Blattman, Jamison and Sheridan (2017) and Karlan and Zinman (2012) document substantial error in small-scale surveys typical of randomized control trials and development studies. For U.S. household surveys, the quality of survey data has been declining steadily in recent years.[1] Households are more reluctant to participate in surveys, and participants are more likely to refuse to answer particular questions and give inaccurate responses. Non-response rates have been increasing for nearly all surveys in the U.S., and a large literature attempts to understand the causes and consequences of this trend (e.g., Groves 2006, Groves 2011, Massey and Tourangeau 2013 and Meyer, Mok, and Sullivan 2015).

Less is known about measurement error, i.e. how the reported responses of households differ from true values, even though its relevance has been documented for many variables, including education (Black, Sanders, and Taylor 2003), drug use (Johnson and Fendrich 2005) and self-reported health status (Butler et al. 1987). These reporting errors have been shown to be predicted by respondent characteristics or the true value of the variable (e.g., Bollinger and David 1997), so they lead to bias in most common analyses (e.g. Bound, Brown, and Mathiowetz 2001; Meyer, Mittag and Goerge, 2021). These biases can be sizeable. For example, Meyer and Mittag (2021a) find that measurement error generates large biases in survey data, with the magnitude of the bias in mean reports being more than three times that due to survey non-response. While the presence and magnitude of errors has recently become widely documented, our understanding of the reasons

[1] See Massey and Tourangeau (2013), Meyer, Mok, and Sullivan (2015) and the more general reviews in e.g. Biemer et al. (1991), Bound, Brown, and Mathiowetz (2001), Alwin (2007), and Groves et al. (2009).

for errors is meagre. As a consequence of this lack of knowledge about the nature of errors and hence about strategies to work with contaminated data, researchers typically at best acknowledge the (likely) presence of errors in their data. One reason for the dearth of empirical research on measurement errors in surveys is that reliable measures of "truth" for survey variables are rare.

In this paper, we study measurement error in surveys and analyze theories of its nature in order to improve the accuracy of survey data and estimates derived from it. In particular, we study measurement error in reports of participation in government programs by linking the surveys to administrative records. We argue that this data linkage can provide the required measure of truth if the data sources and linkage are sufficiently accurate. We work with high quality administrative records of the Food Stamp (SNAP) and Public Assistance (PA, which combines TANF and General Assistance) programs in New York State that are linked to three of the most important U.S. household surveys: the American Community Survey (ACS), the Annual Social and Economic Supplement (ASEC) to the Current Population Survey (CPS), and the Survey of Income and Program Participation (SIPP). Using the linked data, we have measures of program participation for the same observation from survey data (reported measure) and administrative records (true measure). Our linked data thereby put us in a unique position to study misreporting, even compared to previous validation studies. Prior studies often have little generality, using one survey, a single program, or data that is 30 years old.[2] We provide a more powerful examination of the reasons for error by comparing results across different surveys and different programs.

Specifically, we study two types of errors in binary variables: false negative responses (failures of true recipients to report) and false positive responses (reported receipt by those who are not in the administrative data). We first confirm that the extent of survey error is large, with up to 59 percent of non-imputed true recipients failing to report program receipt (in the case of PA in the CPS). As prior studies have documented, a combination of high false negative rates and low false positive rates leads to a substantial net understatement of program receipt (Celhay, Meyer and Mittag, 2021a). Substantial error arises from imputation, but reporting error is by far the largest source of aggregate error and varies substantially between surveys and programs (Meyer and

[2] See David (1962), Marquis and Moore (1990), Bollinger and David (1997), Meyer, Mok, and Sullivan (2015) and Meyer, Mittag, and Goerge (2021). See Bound, Brown, and Mathiowetz (2001) for a review and Bruckmeier, Müller, and Riphahn (2014) for a study of misreporting of welfare receipt in Germany.

Mittag, 2021a). Our main questions in this study are:[3] Why do people misreport after agreeing to answer a question? Are these errors due to cognitive limitations or unwillingness to reveal known answers? We test whether variables that measure important factors in common theories of misreporting explain false positive and false negative errors. Our unusually detailed data allow us to construct measures of key concepts that are hypothesized to affect survey error, including cognitive factors, the willingness of respondents to cooperate and survey design choices.

In terms of cognitive issues, we examine the role of recall error and salience. We show that underreporting increases with time passed between receipt and the interview, i.e. recall errors lead to false negative errors. In addition, failure to recall the timing of receipt leads to overreporting by households who received benefits before the interview, but not during the reference period. These two recall problems account for a sizeable share of reporting errors. We provide evidence that salience reduces reporting error, as households that are more dependent on government transfers are better reporters on average. To assess the role of the willingness of respondents to reveal known answers, we examine stigma and cooperation with the survey overall. We find that households in ZIP codes with higher program participation rates are more likely to reveal true participation, suggesting that stigma leads to underreporting of "socially undesirable behavior". We also find that households that are less cooperative with other sections of the survey, as measured by the fraction of questions they refuse to answer, are more likely to give incorrect answers. Finally, we examine the effects of several survey design choices. We find that proxy interviews are as accurate as other responses. Our results on survey mode are in line with the previous literature, which suggests a trade-off between non-response and survey accuracy. However, assignment to mode is not random, so these findings could reflect who is interviewed by what mode. Yet, repeating our analyses by survey mode provides evidence of the validity of our results: The effects we document are present across modes, with the exception of the effect of stigma, which, as expected, only affects response error in the presence of interviewers.

[3] We have used the same linked data to examine related questions on the extent of survey error (Celhay, Meyer and Mittag, 2021a), the sources and extent of bias it causes (Meyer and Mittag 2019a, Meyer and Mittag, 2021a, Celhay, Meyer and Mittag 2021b) and how linked data can be used to correct estimates (Davern, Meyer and Mittag, 2019; Mittag, 2019).

This paper contributes to a large literature that uses survey data to study poverty, the income distribution, and program participation and its effects,[4] by analyzing errors in reported receipt of government benefits. Our findings are also related to a large literature that analyzes data quality in household surveys.[5] We use high quality validation data, multiple surveys, and programs to provide extensive evidence of what causes measurement error. Our findings have many implications for both data users and data producers that are of importance beyond the case of government transfers and the specific surveys that we study. Our results are also important for the large number of researchers in economics who conduct their own surveys. The patterns of misreporting we document allow data users to gauge the prevalence of errors in their data and select the most reliable measures. We argue that survey respondents behave like economic agents whose survey responses are shaped by the costs and benefits of providing truthful survey responses. This view, which is supported by our empirical findings, can help researchers assess likely sources of errors even in questions very different from benefit receipt. Understanding which variables reliably predict errors can help researchers devise strategies to reduce or correct bias and to avoid corrections that are unlikely to work. Survey producers can improve survey data by making predictors of error available to survey users, who could use them to measure data accuracy and to correct estimates using, for example, instrumental variable strategies. By clarifying the conditions under which misreporting arises, our results can also guide efforts to reduce survey error. Our findings particularly highlight two trade-offs that survey designers need to balance according to their objectives: increasing response rates likely decreases response accuracy and asking for more detail may also result in more error. Our results are broad enough to be applied to many examples where researchers believe that measurement error from misreporting is a problem, including studies of health, crime, or earnings.

This paper is structured as follows. In section II we summarize theories of misreporting. In section III we describe the data and how we link administrative records to survey data. In section

[4] See e.g. Fraker and Moffitt (1988), Blank and Ruggles (1996), Deaton (1997), Gleason, Schochet, and Moffitt (1998), Currie et al. (2001) , Gittleman (2001), Gundersen and Oliveira (2001), Blank (2002), Grogger (2002), Danielson and Klerman (2006), Almond, Hoynes, and Schanzenbach (2010), Hoynes and Schanzenbach (2012), Moffitt (2014) and Meyer and Wu (2018).

[5] Some examples of this literature are Sudman and Bradburn (1974), Marquis and Moore (1990), Bollinger and David (1997), Bound, Brown, and Mathiowetz (2001), and Groves et al. (2009).

IV we test theories of misreporting. In section V we discuss how our results can help survey users and producers reduce the problem of measurement error. Section VI concludes.

II. Reasons for Errors in Survey Data

This section briefly reviews the literature on reasons for misreporting in surveys. To organize the discussion, we divide the theories of misreporting, as do Mathiowetz, Brown, and Bound (2002) into survey design, and theories related to respondents' characteristics and behavior. Our review focuses on sources of error on which we provide evidence below. See Alwin (2007) and Groves et al. (2009) for reviews of the literature on survey design and Sudman and Bradburn (1974), Sirken (1999) and Bound, Brown, and Mathiowetz (2001) for reviews of the literature on respondents' characteristics and behavior. We further divide the latter category into errors arising from cognitive issues and cooperation.[6] Cognitive factors can lead to errors due to respondents' inability to provide a correct answer, e.g. because they are unable to recall it. A lack of cooperation can lead to errors due to respondents' unwillingness to reveal the true answer even though they know it, e.g. due to stigma. From an economic perspective, a respondent should provide a correct answer if revealing the truth yields higher utility, i.e. benefit exceed costs, than an incorrect or no answer. A respondent has to be willing to incur the non-monetary cost of the effort to comprehend the question, to attempt to recall the correct information or look it up somewhere, and finally to communicate it to the survey producer.

A. Respondent characteristics and behavior: Cognition

Errors due to cognitive issues stem from the cognitive effort to comprehend the question and (accurately) retrieve the required information. For a detailed discussion on the cognitive aspects of survey response, see Tourangeau (1984), Tourangeau and Rasinski (1988), and Tourangeau, Rips, and Rasinski (2000). We study three issues related to cognitive factors: recall of the event, telescoping of events, and salience.

There is evidence that the effort required to retrieve information and hence the accuracy of responses depend on how the survey uses reference periods in questions (Groves et al. 2009, p. 231). According to Sudman and Bradburn (1973), memory errors in surveys can take the form of recall errors, in which the respondent forgets the episode entirely, or telescoping errors, in which

[6] The literature provides several more nuanced classifications, see e.g. Groves et al. 2009, pp. 218-224.

respondents mistakenly place events in more recent (forward telescoping) or previous periods (backward telescoping). An effect of length of recall on underreporting in surveys has been found in studies on consumer expenditures (Neter and Waksberg 1964) and self reports of drug use (Bachman and O'Malley 1981), among others. Consequently, shorter and more recent reference periods may lead to more accurate answers through lower recall costs. Gray (1955) conducted an early experimental study of telescoping, showing that employees can accurately remember the type of sick leave they took but not the dates in which they did so. See e.g. Gaskell, Wright and O'Muircheartaigh (2000) for a more recent study. Finally, the effort required to retrieve accurate information may be lower if the event is more salient, i.e. more frequent or highly present in a respondents' mind. Therefore, respondents to whom the topic in question is more important or on whose mind it is more frequently have been argued to give better answers (Sharp and Adua 2010).

B. Respondent characteristics and behavior: Cooperation

The second group of error sources arises from respondent cooperation, i.e. whether respondents are willing to devote the effort to provide accurate responses. Even if the answer is known, so that recall costs are low (or sunk), the respondent also has to be willing to reveal the answer to a stranger, potentially incurring further costs such as social stigma if the topic is sensitive (Karlan and Zinman 2008). A frequently discussed cost of reporting true behavior is the cost of providing socially undesirable answers (DeMaio, 1984; Tourangeau and Yan, 2007). This type of stigma has been shown to be related to inaccurate answers to sensitive questions such as abortion (Fu et al. 1998) and drug use (Brittingham, Tourangeau, and Kay 1998), for example. Likewise, social desirability can also lead to overreporting of socially desirable behavior such as voting (Belli, Traugott, and Beckman 2001). In the case of government transfers, some argue that stigma may explain why eligible individuals do not participate in government programs such as cash welfare or SNAP (Moffitt 1983). Similarly, stigma could also explain why people do not report participation to interviewers, leading to underreporting.

More generally, subjective benefits and costs of providing truthful answers may determine the cooperativeness of respondents with a survey overall. According to Krosnick (1991), the quality of survey data is also affected by respondents' motivation or attitudes towards surveys. Respondents who are less willing to cooperate with a survey may well engage in "suboptimal response strategies", such as "satisficing". Evidence on the relation between cooperativeness and response quality is scarce and inconclusive. Bollinger and David (2001) provide evidence that

more cooperative respondents are less likely to misreport program receipt in the SIPP. In another study, Kaminska, McCutcheon, and Billiet (2010) show that reluctant participants in surveys give low quality responses. However, the relationship disappears when they control for cognitive measures. Understanding the relationship between cooperativeness and survey quality is important, since efforts to increase response rates may include less cooperative participants and thereby increase error rates.

C. Survey features

The literature on "Total Survey Error" provides a detailed classification of components of survey design that can affect data quality (see Groves and Lyberg 2010). An important design question in household surveys is which and how many members of a given household to interview. Interviews in which the respondent provides information about other individuals are called proxy interviews. The evidence on the effect of proxy interviews on survey accuracy is mixed. They can improve over self-reports (Tourangeau, Rips, and Rasinski 2000, Bollinger and David, 2001), increase error rates (Cartwright, 1957, Tamborini and Kim 2013), or may not differ from self-reports (Moore 1998). Consequently, the literature has not determined whether proxy interviews are a cause of or a remedy for survey errors.

The choice of survey mode, i.e. whether responses are obtained from face-to-face interviews, interviews assisted by computer software (CAPI), telephone interviews (CATI), or by self-administered mail-back questionnaires, may affect both the decision to participate in a survey and the response quality of participants (Kanuk and Berenson 1975; Lyberg and Kasprzyk 1991). In a meta-analysis, de Leeuw (1992) finds that response quality varies significantly across interview modes. The evidence points to a trade-off between response rates, which are lower for self-administered questionnaires, and response accuracy, which is lower for modes that require interviewers. The magnitude and direction of these effects depend on the context. Participation in government programs is usually considered a sensitive topic, suggesting that interview modes with higher privacy to the respondent, such as self-administered surveys, will have lower misreporting rates. On the other hand, interviewers may also increase response quality (e.g. Bruckmeier, Müller, and Riphahn 2015), for example by facilitating comprehension and recall or by double-checking on answers that are likely to be wrong. So while survey mode and interviewers may have important effects on survey quality, the evidence regarding how they affect response quality and the conditions under which they do is inconclusive.

III. Data

The theories summarized in the previous section are about errors at the level of the unit of observation. Consequently, one needs a measure of "truth" at the unit level to study them. An additional difficulty in assessing theories of survey errors is that important determinants of errors, such as the survey design, do not vary within survey. We argue that linking survey data to administrative records can solve the first problem. We link administrative records on two transfer programs to three major household surveys to mitigate the second problem.

A. Survey data

We study misreporting of program receipt in the ACS, CPS and the SIPP. Each survey contains demographic information, receipt of government assistance, labor force participation, education, and other variables. The specific questions that we use can be found in Table 1.

The ACS samples approximately 2.5 percent of the U.S. population each year. It is the largest U.S. household survey. Interviews are spread across all months of the year, with more than 290,000 households selected each month to participate. The ACS questionnaire is similar to the long-form decennial census and is administered by mail, telephone, or face-to-face interview. In terms of information on government transfers, it asks for participation in SNAP but not for the amount received. It asks for both receipt and amounts received for PA. For both programs, the questions in the ACS refer to the 12 months prior to the interview date. We use the ACS for survey years 2008 through 2012.

The CPS is one of the most important economic surveys 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 ASEC supplement of the CPS, which also is the official source of income and poverty statistics in the U.S. The ASEC is conducted in February, March, and April during the 2008-2013 interview years that we use. The CPS asks for participation and total dollars received from SNAP and PA during the previous calendar year, 2007-2012 in our case.

Finally, the SIPP is the highest quality source of information on poor households and receipt of government transfers. We use the 2004 panel (waves 10 through 12) and 2008 panel (all waves) of the SIPP that each consist of approximately 50,000 households who are followed for a period of 4 years. The survey provides monthly information on participation and dollars received from most government transfer programs in the U.S., including SNAP and PA. Like Ribar (2005) and

Acs, Phillips, and Nelson (2005), we aggregate program participation and total amounts received over a four month wave for each household and analyze each wave as a separate cross section.[7]

B. Administrative data and data linkage

We link the three surveys to administrative records on SNAP and PA benefits from the Office of Temporary and Disability Assistance of the State of New York (NY OTDA). The data contain information on the universe of monthly payments for SNAP, Temporary Assistance for Needy Families (TANF), and General Assistance in New York State from January 2007 through December 2012. [8] We aggregate TANF and General Assistance to PA to abstract from errors due to confusing these two programs. Each record in the data corresponds to a monthly payment to a specific case and includes information about geographical location, number of members in each case, birth date of each member, and other demographic characteristics. As part of eligibility determination, applicant information is checked by OTDA against Social Security records. The records are from actual payments, which are audited, and appear to be accurate. For SNAP, the overall total dollars from our administrative records differs from official aggregate outlays by less than a percent in all years.[9] These data have been previously used by Meyer and Mittag (2019a, 2021a) and Celhay, Meyer and Mittag (2021a), who further discuss their accuracy.

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.[10] 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 derived from the Social Security Administration Numerical Identification file. The reference file 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)[11] and attached to the corresponding records in our data. A PIK is obtained

[7] To account for the dependent sampling, we cluster standard errors at the household level in the SIPP.
[8] In the years since the 1996 welfare reform act, General Assistance has grown relative to federal cash assistance. In recent years, total benefit payments have exceeded those of TANF in New York. Likewise, SNAP has experienced a large increase in its caseload in recent years, making it one of the largest in-kind transfer programs.
[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.
[10] See Wagner and Layne (2014) on how administrative data and surveys are linked at the US Census Bureau. See Ridder and Moffitt (2007) and Chun et al. (2021) for reviews of linking administrative records to surveys.
[11] PIKs are anonymized Social Security numbers used to protect the identity of respondents.

for over 99 percent of the administrative records from each program. Our unit of analysis is a household, which is logical given the sharing of resources among members. Using households also ensures a high rate of data linkage. Since the administrative data include records for each recipient, we are able to link the information from a program case to the household if any true recipient in the household has a PIK. Therefore, we consider a household to have a PIK if a PIK was obtained for anyone in the household.[12] The PIK rates at the household level are 93 percent in the ACS, 91 percent in the CPS, and 95 percent in the SIPP. In order to account for incomplete linking, we multiply the household weights by the inverse of the predicted probability of any household member having a PIK (Wooldridge 2007).[13] As the high rate of PIK-linking suggests, our results do not appreciably change when using the unadjusted household weights.

C. Linked Data and Extent of Error

We argue that the linked data provides a sufficiently error-free measure of receipt to study survey error at the household level. This unusual accuracy of our data stems from using high quality administrative data and achieving a high match rate.[14] Our administrative measure stems from validated payments, which makes them very accurate even compared to other administrative records such as income obtained from IRS records, where both the survey and the administrative measure may contain errors. Working with the universe of SNAP and PA cases provides us with an accurate measure of truth for both recipients and non-recipients. Unlike studies that link survey data to a sample of recipients, linking them to the universe of recipients allows us to study both failure to report receipt and overreporting of receipt. More than 99 percent of our administrative records have a PIK, so that any substantial bias from imperfect linking comes from survey observations for which a PIK is not available. The PIK rate in all three surveys is high, so errors due to imperfect linkage should be rare.

---

[12] 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.

[13] The coefficients of the probit model we use to predict these probabilities are available upon request.

[14] We do not mean to argue that administrative or linked data are more accurate in general. Administrative data can have substantial error, see Niehaus and Sukhtankar (2013) for an extreme case. See Courtemanche, Denteh and Tchernis (2019) and Meyer and Mittag (2019b) for a discussion of a specific linked data source. See Meyer and Mittag (2021b) for further discussion.

The three surveys ask about program receipt at any point in the reference period, giving us a measure of reported participation. Our linked administrative data allow us to match the definition of program participation in the survey data, which provides us with a measure of actual receipt during the reference period of the survey. Using these two measures, we define two types of errors: False negatives arise from households that actually receive a program, but fail to report receipt when asked in the survey. False positives arise from households that did not receive benefits from a program during the reference period of the survey, but were recorded as having received aid from that program in the survey. Celhay, Meyer and Mittag (2021a) show that imputed observations account for a large share of false positive errors overall, even though only a small share of observations are imputed. In the remaining analyses, we exclude imputed observations in order to study errors from misreporting separately from imputation errors.

Figure 1 summarizes the aggregate error rates in our sample.[15] For example, the false negative rate in the CPS is 59 percent for PA, i.e. almost six out of ten PA recipients are recorded as non-recipients. In addition, 0.03 percent of households that are not true recipients are recorded as having received PA in the CPS. The false negative rate is much higher than the false positive rate in all three surveys, leading to substantial net underreporting (Meyer, Mok and Sullivan, 2015).

The accuracy, size, and detail of our data are unusual, even compared to other studies that use validation data to assess measurement error in household surveys. Studies that use administrative microdata linked to surveys typically use data covering a short time period, from one survey, one program, and/or a small subsample of respondents. As such, any conclusion from these studies may be particular to the survey or sample used and the period of analysis, limiting the generalizability of their results. With our data, we are able to study measurement error at the household level using six years of data and compare results across two programs and three surveys.

Four features of our data are particularly important. First, the high match rate not only makes our data unusually accurate, but also allows us to work with large samples, using more than 90 percent of the sample in each of the surveys. Even after excluding imputed values, our final sample contains 15,207 households in the CPS, 448,135 households in the ACS, and 20,434 household-wave observations in the SIPP.[16] These are large samples in comparison to previous studies, which

[15] See Meyer and Mittag (2021a) and Celhay, Meyer and Mittag (2021a) for detailed analyses of the extent of error in these data.

[16] These sample sizes refer to our analyses of SNAP, the sample sizes for PA differ slightly due to the differences in imputation rates: 15,476 (CPS), 415,656 (ACS) and 20,277 (SIPP).

were often limited in the theories they could test and their statistical precision to discriminate among theories by their small sample sizes.[17] Second, our data are unusually detailed, since we start with household surveys that contain rich information. Linking the surveys to the longitudinal information from the administrative records allows us to know whether a household received SNAP or PA in any month during calendar years 2007 to 2012, regardless of the timing and length of the reference periods in questions about program receipt across the three surveys. Taken together, this detail allows us to examine multiple theories of the causes of survey error in the same data. Third, we are able to compare errors across three surveys and across two programs in each survey. These multiple comparisons allow us to better distinguish reliable determinants of misreporting from spurious or case-specific results. Comparing errors across surveys and programs also provides an unusually rich analysis of the determinants of errors. Since each survey is a random sample of the same population, one might expect similar error rates. However, Figure 1 shows important differences across surveys, suggesting that survey design can substantially affect error rates. In fact, the designs of the three surveys we use differ in many dimensions. Linking them to the same administrative records with the same linkage methods allows us to assess whether differences in survey design affect survey accuracy. Within each survey, the error rates are also different across programs, suggesting that characteristics of the program or its participant pool may also affect reporting accuracy. Our data allows us to analyze such conjectures by comparing error rates across the two programs within the same survey using the same linkage method, while also holding constant the survey design. Finally, using calendar years 2007 to 2012 enables us to examine measurement error in a more recent period than other studies and at a time when the SNAP caseload was growing rapidly.

IV. The Nature of Errors in Reported Program Participation

To understand the determinants of measurement errors, we examine how survey design and respondent behavior affect errors in our data. To do so, we estimate Probit models that include variables related to theories of response error. For each program and survey, we estimate separate Probit models for false positives and false negatives, because the two error types may be explained by different factors. For the false negative Probits, we restrict the sample to true recipients and

[17] See Marquis and Moore (1990), Bollinger and David (1997), and Meyer, Mittag and Goerge (2021).

estimate the determinants of the probability that they fail to report receipt. Similarly, for the false positive Probits, we estimate the probability of mistakenly reporting receipt among true non-recipients. Even after excluding imputations, false positives may still be considered less informative about misreporting than false negatives, as they may also arise from other forms of data editing or the linkage process.[18]

We include demographic controls in all models: number of adults and children, sex, age, education, race, disability, household income, citizenship status of the household head and whether (s)he speaks English poorly, whether the household is in a rural area, reported receipt of other programs, and a linear trend in calendar years. Observations are weighted using survey weights adjusted for the predicted linking probability using Inverse Probability Weighting. Table 2 presents descriptive statistics for each variable we analyze for SNAP and Table 3 for PA.

As section II discusses, the literature has proposed several explanations for why a substantial share of survey respondents agrees to answer a question but then provides an erroneous answer. Our main variables of interest measure factors that these theories hypothesize to be related to survey error. We start by examining cognitive factors (recall, mistimed reports and salience) and then turn to measures of respondents' willingness to reveal known answers (stigma and cooperation). Finally, we examine how proxy interviews, survey mode and the presence of interviewers shape the extent of survey error. We explain how we construct each variable as we describe the results below. All specifications include all variables at once as well as demographic controls. Thus, the results are partial effects holding other determinants of misreporting fixed.[19]

A. Respondent characteristics and behavior: Cognition

The first cognitive factor we examine is recall, in particular the ability of the respondent to recall receipt of the program and whether this receipt was during the reference period of the survey. The administrative data record monthly receipt, so we can test whether the number of months gone

[18] For example, as we discuss in Meyer and Mittag (2021a) households that moved from another state recently may truly report participation in that state so that they are not really false positives, but we recorded them as such, because we only observed whether they received any aid in New York State during the reference period. See Meyer, Mittag and Goerge (2021) for a discussion of the consequences of errors arising from data linkage.

[19] Whether the coefficient estimates can be given a causal interpretation requires further discussion and varies across coefficients. Our rich set of controls for demographic factors and other determinants of misreporting suggest that the effects we find provide evidence on the causal relationships postulated by the theories we test. But the strength of the evidence varies between our measures. For example, recall periods are mainly determined by the timing of the survey interview relative to last receipt, making the exogeneity assumption plausible, but the case is less clear for our measures of salience.

by between the last transfer and the interview affects the rate of underreporting. This analysis is done holding constant the number of months households participate in each program, obtained from the administrative records. This strategy makes it more plausible that the length of the recall period is exogenous, since it mainly varies due to the timing of the interview relative to last receipt.

The results in Table 4 are remarkably consistent across surveys and programs, particularly for the (more precisely estimated) effects in the CPS and ACS, which both suggest that an additional month since receipt is associated with 3-4 percent of true recipients forgetting to report receipt. Specifically, for both SNAP and PA, true recipient households in the CPS are 3.6 percentage points more likely to fail to report receipt per month passed. This corresponds to a 10 percent increase in the probability of a false negative for SNAP and a 6 percent increase for PA. In the ACS, each additional month since the last receipt increases the probability of a false negative by 3.3 percentage points for SNAP and 3.0 percentage points for PA, corresponding to 13 percent of the average false negative rate for SNAP and 5 percent for PA. In the SIPP, the false negative rate of PA increases by 9.1 percentage points (20 percent of the sample average) while the effect on reporting SNAP is not significant at 2.2 percentage points per month gone by. The magnitude of these effects shows that recall is an important source of survey error and accounts for a sizeable share of the difference in accuracy between our three surveys: A simple calculation of mean differences in the recall period shows that if the CPS had the same recall period as the ACS, the false negative rate for SNAP in the CPS would go down by 7.1 percentage points or 57 percent of the difference between the error rates of the two surveys.[20] Similarly, aligning recall periods would reduce the difference between the CPS and the SIPP by 46 percent.

Respondents may still misreport if they correctly recall program receipt, but do not remember the time of receipt correctly. They may mistakenly bring forward receipt before the reference period (telescoping) or report receipt between the reference period and the interview (backward telescoping). In the CPS, for example, interviews are between February and April, but the reference period is the previous year. Telescoping may lead to false positives as recent events may be pushed forward or backward into the reference period. To test this hypothesis, we construct

[20] The average recall period is 3.52 months in the CPS and 1.56 months in the ACS. Multiplying their difference by the marginal effect of months since last receipt in the CPS (0.036) implies a 7.1 percentage point reduction in the false negative rate. The average recall period in the SIPP is 1.09 months, implying a difference of 2.44 months to the CPS and hence a reduction of the false negative rate by 9 percentage points (46 percent of the 8.8 percentage point difference in the error rates).

an indicator that equals one if a household that did not receive a transfer within the reference period participated in the program before the start of the reference period according to our administrative data. For the CPS and the ACS, we use twelve months before the start of the reference period, and in the SIPP we use the last four months before the start of the reference period to match the length of the reference periods of the surveys. The CPS additionally allows us to test for a backward telescoping effect by defining a binary indicator that equals one if a household received a transfer after, but not in, the reference period and before the interview month, i.e. between January and April of the year of the interview.

The results in Table 5 provide evidence that telescoping may partly explain false positive errors, but some of the estimates are imprecise. In the ACS, we find that receipt of SNAP before the reference period increases the likelihood of a false positive response by 0.7 percentage points. For PA, this telescoping effect increases the probability of a false positive response by 1.2 percentage points. For both programs, the telescoping effect corresponds to more than 60 percent of the false positive rate. In the SIPP, receipt before the reference period increases the likelihood of mistakenly reporting receipt by 1.2 percentage points for SNAP and 1.1 percentage points for PA. Again, the telescoping effects are sizable, as they more than double the probability of a false positive response for SNAP and almost quadruple it for PA. For the CPS, the point estimates are of similar magnitude as in the other surveys and large with respect to the average false positive rate, but insignificant. The CPS results on backward telescoping show that receipt after the reference period increases the likelihood of a false positive response for PA by 0.6 percentage points, almost doubling the probability with respect to the sample average. The effect of backward telescoping for SNAP is of similar magnitude, but not statistically significant.

Another cognitive aspect that could affect response quality is the importance or salience of the issue. We construct two measures of salience, the number of months a household received program benefits and the average monthly amount they received during the reference period.[21] Holding constant the number of months gone by since the last transfer, program receipt should be more salient for households with longer receipt or higher amounts, possibly because they are more

[21] Months of participation and monthly amount received are from the administrative data so that they are actual and not reported. We only observe them for true recipients, so we do not test whether salience affects the false positive rate. Both months of receipt and amounts depend on factors such as the severity of need that may be related to reporting accuracy through additional channels, so the case for exogeneity is less clear here than for our analyses of recall.

dependent on government transfers. Recall should be easier for them and therefore, they should be less likely to make mistakes when reporting.

The results in Table 4 confirm that the frequency of false negatives indeed decreases with our measures of salience. In the CPS and ACS, the false negative rates of the two programs decrease by 1 to 2 percentage points for each additional month of receipt in the past calendar year. In the SIPP, an additional month of SNAP receipt during the four-month reference period leads to a 5-percentage point decrease in the probability that receipt is not reported. PA spells in the SIPP that last an additional month are almost 3 percentage points less likely to be misreported, but this effect is not statistically significant. The next row in Table 4 shows the effects of transfer amounts on the probability of a false negative error. In most cases, higher amounts of receipt have a negative effect on the probability of not reporting. An additional $100 of monthly benefits leads to a reduction in the false negative rate by 0.3 percentage points for SNAP in the ACS and a 1.1 percentage point reduction for SNAP in the SIPP as well as for PA in the ACS. The effect is not statistically significant, but suggests a reduction in error, for SNAP in the CPS and PA in the ACS. However, our results show increased error for PA in the SIPP, where the probability of a false negative response increases by 1.5 percentage point with each additional $100 USD received. Overall, our results show that more salient events, as measured by the proxies we constructed, are reported better but the effects are small when compared to other determinants of false negative responses.

B. Respondent characteristics and behavior: Cooperation

Respondents may misreport even when they know the true answer, i.e. when there are no recall costs to giving a correct answer. For government transfers, stigma or social desirability has been discussed as one of the main reasons why individuals fail to truthfully report program receipt. Stigma can be thought of as reducing the benefit or increasing the cost of providing a correct answer, for example through negative judgment by other individuals such as the interviewer. We argue that welfare stigma should be lower in areas with higher rates of participation: Where participation is more common, it should be more acceptable and therefore more likely to be revealed to a third person (but see Besley and Coate, 1992, for a contrary argument). Our measure of local participation is the participation rate of each program in the ZIP code of residence. For each ZIP code in New York, we calculate the participation rate as the ratio of total program cases according to the administrative data to total housing units according to the 2010 US Census. We

then assign these annual participation rates to households in each sample by ZIP code and year. Thus, we test whether stigma increases the probability of false negatives using the proportion of households that participate in each program in the respondent's ZIP code as our measure of stigma. Local participation rates may be correlated with other predictors of misreporting, which would bias our results here. Celhay, Meyer and Mittag (2021b) extend this analysis to provide evidence of a causal effect of stigma.

Table 4 provides evidence that stigma leads to false negatives, but the effects are small. In the CPS, a 10 percentage point increase in local SNAP participation is associated with a decrease of 0.8 percentage points in the probability of not reporting true program participation. For PA, a 10 percentage point increase in the local PA participation rate is associated with a decrease of 3.2 percentage points in the probability of failure to report true participation. In the ACS, a 10 percentage-point increase in local PA participation reduces the likelihood of underreporting PA by 1.3 percentage points. We do not find a significant effect for SNAP in the ACS. In the SIPP, the point estimates are large and positive, but imprecise. Overall, the results provide evidence that stigma matters in program participation behavior, in this case the report of program participation to a third party. However, the effects are small relative to other factors that affect underreporting.

Finally, we study how willingness to cooperate with the survey overall relates to misreporting. We measure cooperativeness using the fraction of other survey questions a household refused to answer, i.e. the fraction of responses that are imputed. Zabel (1998) finds that a similar measure based on imputation counts predicts attrition in the SIPP and the Panel Study of Income Dynamics, which suggests that item non-response is related to cooperation with the survey overall. Hence, households that refuse to answer a larger share of question are characterized as less cooperative.[22] The imputation procedures and frequencies differ between surveys, so we standardize our measure of cooperation: We include two indicators in our Probit models for the location of a household in the distribution of imputation rates within a survey: one for a household being between the 75th and 90th percentile (indicating low cooperation), and one for a household being above the 90th percentile (indicating very low cooperation). Households below the 75th percentile (cooperative households) are the omitted base group.[23] Following Bollinger and David

[22] We select questions that are posed to every household in each survey, see Appendix Table 1 for a list.
[23] Note that the distribution of the frequency of non-response is standardized before we remove imputed observations, which explains why the sample proportions do not correspond to the percentile categories exactly.

(2001), we also use the longitudinal data from the SIPP to test whether households that leave the sample in future waves of the survey (attrition) report less accurately. Our results show that less cooperative households indeed misreport more frequently.

For false negatives, Table 4 shows that errors tend to increase as survey cooperation decreases. In the CPS, recipient households with low cooperation are 5.6 percentage points more likely to fail to report receipt of SNAP, while households with very low cooperation are 31 percentage points more likely to do so. Thus, the probability of misreporting SNAP receipt increases by 15 percent of the average false negative rate for a low cooperation household . For very low cooperation households, the probability of failing to report SNAP receipt increases by 82 percent of the average false negative rate. The effect on false negative responses for PA is small and not significant in the CPS. The results for the ACS also show that measurement errors become more frequent with lower cooperation, but the effects are smaller. The probability that a recipient household with low cooperation fails to report receipt in the ACS is 1 percentage point higher for SNAP and 2.5 percentage points higher for PA, all else equal. For both programs, the increase corresponds to 4 percent of the false negative rate. Households with very low cooperation are almost 4 percentage points more likely to fail to report true receipt of either program in the ACS. However, the results for the SIPP are not significant. We also do not find an effect of future attrition in the SIPP.

The results in Table 5 show that our measure of survey cooperation also predicts false positive responses. For SNAP in the CPS, low cooperation increases the probability of a false positive response by 38 percent of the overall false positive rate. The point estimates are similar, but insignificant, for households with very low cooperation. For PA, our estimates imply an increase in the probability of a false positive response by 117 percent of the average false positive rate for low cooperation households. In the ACS, the probability of a false positive SNAP response increases by 10 percent of the average false positive rate for low cooperation households and by 29 percent for very low cooperation households. For PA, the effect for low cooperation households is small and insignificant in the ACS. However, the probability of a false positive response is 35 percent of the overall false positive rate higher for very low cooperation households. In the SIPP, the probability of a false positive response for PA is higher by 93 percent of the overall false positive rate for households with low cooperation. The other effects, including the effect of future attrition, are not significant in the SIPP.

C. Survey design: Proxy Responses, Interview Mode, and, Presence of Interviewers

Finally, we examine the relation between survey design features and survey error. We first use the SIPP to examine the accuracy of proxy responses, i.e. responses provided by another household member for someone who declined answering the survey. Proxy responses have been argued to be less accurate due to less information (Cartwright, 1957, Tamborini and Kim 2013) or more accurate due to higher cooperativeness (Tourangeau, Rips, and Rasinski 2000, Bollinger and David, 2001). We do not find proxy responses to differ significantly in accuracy from other responses, indicating that the two effects are either small or cancel. This finding suggests that survey producers may be able to reduce non-response without reducing data accuracy by allowing for proxy responses.

We next examine the relation between interview mode and misreporting by including indicators for survey mode in the Probits for the two error types. Of our ACS sample, 38.2 percent are interviewed in-person, 7.8 percent are interviewed by telephone and the remainder is self-administered by mail. In the CPS, 12.7 percent of our sample is interviewed by telephone while the remainder is interviewed in-person. We are not able to explore interview mode effects in the SIPP since it only conducts in-person interviews. Unfortunately, respondents are not assigned to modes randomly. [24] In both the CPS and ACS, it is reasonable to conjecture that households interviewed by phone are more accurate respondents than those interviewed in person, because they have been more cooperative with previous contact attempts. Those who respond to the mail-back ACS questionnaire may well be even better respondents through self-selection, since they have responded to the first contact attempt. This sample selection may bias our estimates of mode effects, so they should be interpreted with caution.

In the ACS false negative rates for SNAP are higher for telephone interviews compared to mail-back questionnaires (by 8.1 percentage points) and even higher for face-to-face interviews (by 16.4 percentage points). Similarly, for PA, telephone interviews are 6.6 percentage points more likely to yield a false negative response compared to mailed responses, while in-person interviews

[24] The ACS sends a questionnaire by mail to every sampled household (U.S. Census Bureau 2014). A sample of households that fail to send it back are contacted by telephone. If telephone interviews cannot be conducted, households are visited for an in-person interview. In the CPS, households are interviewed by telephone rather than in-person if i) households have a telephone and accept a telephone interview, ii) the field representative recommends a telephone interview, and iii) the interview month is neither the first nor the fifth interview of the household (US Census Bureau 2006).

are 14.3 percentage points more likely to do so. Likewise, in the CPS the false negative rate for SNAP is 15.3 percentage points higher in face-to-face interviews than in telephone interviews. For PA, in-person interviews are not significantly more likely to yield false negatives. This pattern is consistent with prior research on survey mode as well as the self-selection we describe above, so these results do not allow us to tell these mechanisms apart.

For false positives, selection and survey mode effects should work in opposite directions. Selection should still increase error rates from mail to telephone to in-person interview as argued above. However, according to the prior literature focused on stigma, the involvement of interviewers should decrease program reports and hence false positive rates. Our findings in Table 5 are mixed, but suggest that mode effects are present: The results for PA in the ACS provide evidence of a mode effect, since the false positive rate is a percentage point lower in telephone surveys and face-to-face interviews when compared to surveys responded by mail. However, the remaining results could also be driven by selection only.

The large sample size of the ACS enables us to estimate the same models as above separately for the mail and non-mail subsamples. The results in Table 6 show that most substantive conclusions hold within survey mode. The effects of months elapsed and salience are very similar across modes of interview. The relation between our measures of cooperation and the error rates is more mixed, as one would expect, because cooperation varies across modes as we discuss above. However, with the exception of false positive responses of PA for the non-mail sample, our results confirm that less cooperative households are more likely to misreport. The results regarding telescoping and stigma hint at more complex interviewer effects. While telescoping effects for SNAP are very similar across modes, the telescoping effect for PA almost doubles in self-administered surveys compared to phone or in-person interviews. The differences across survey modes may reflect that interviewers are useful in clarifying questions, for example regarding the reference period or which government program(s) the question refers to. Finally, our measure of stigma has a negative and significant effect only when respondents are interacting with interviewers for both SNAP and PA. This is consistent with the idea that stigma should only matter when program participants are revealing participation to another person.

These results by survey mode are reassuring, as they provide evidence that our results are not driven by the non-random assignment to survey mode. That we find no effect of stigma without interviewers, while all other results remain as they should, further corroborates our findings above.

Substantively, the differences provide further suggestive evidence that interviewers are important for survey error (Meyer and Mittag, 2019b), but affect survey errors in complex ways.

V. Discussion and Implications

Our results provide new evidence on the reasons for the high rates of survey errors that previous studies reveal, which can help data users make better use of the contaminated data and survey producers improve the data.

For data users, our results are informative about the likely severity and nature of survey errors in the data they use and can help them to improve the way they use the data. The patterns of misreporting we document can help them assess data accuracy in various ways. First, researchers often have some leeway in the choice of the variables they study and our results can help them decide which variables are more accurate. On grounds of such considerations, researchers often choose variables they consider to be less affected by stigma, but our results emphasize the importance of recall as an important source of error. For example, studies that are interested in whether a household receives a program are likely to be less problematic than those studying the particular timing of enrollment and exit. Our results suggest that studies of the determinants of program take-up that look at current receipt have to be less concerned about misreporting than those examining take-up over a longer period, since people tend to forget or confuse the length of their program participation spells. In a similar vein, we show that salient events are more likely to be accurately reported. This finding suggests that short spells are less likely to be reported more generally, which should be taken into account when studying topics such as employment dynamics. On the other hand, it is reassuring for analyses of changes, such as studies of program receipt over a recession period, that the effects of salience are not as large as recall effects. Since the salience of program receipt is likely to change over the business cycle (either through media coverage or changes in program income), strong salience effects would bias the estimated response of government programs to economic conditions. The small salience effects we find are thus in line with the evidence that surveys capture trends better than levels. Our results can thereby help researchers to pick the most reliable variable that is available as well as to understand likely limitations to its accuracy.

Our results also allow researchers to assess the likely extent of error in their data. They can do so informally by considering whether important sources of error, such as failing to recall an

event or its timing, are likely to be relevant. Our findings support the idea that the costs and benefits of providing accurate answers determine whether survey respondents commit errors. This view of survey respondents as economic agents can help survey users assess which sources of error are likely to matter beyond the case of benefit receipt. A more formal approach could make use of the variables that we establish as predictors of misreporting, such as our measures of cooperation. Researchers could take similar questions from the survey they use to estimate error in their sample. For surveys and variables that differ substantially from our study, it seems likely that the relationship between the predictors and errors differs from the one we find here. Such differences make it difficult to predict the extent of error accurately. Yet as long as these measures predict survey error, researchers can still use them to examine likely consequences of error in their data. Computing predictions of accuracy such as our non-response-based measure of cooperation can identify subsamples that are more reliable. Comparing the results from using a subsample of more reliable reporters to the overall sample may allow researchers to assess the effects of misreporting and the robustness of their conclusions.[25]

Studies that compare subpopulations, such as receipt rates by demographic group, may also find it useful to examine whether these groups differ substantially in their response accuracy according to our predictors of misreporting. Knowing which subpopulations are likely to report accurately is important to gauge the populations for which the survey yields reliable statistics and which comparisons are valid. Comparisons between subpopulations always compound differences in true outcomes and differences in reporting, so this strategy can help to provide an impression of the likely severity of this bias and hence which comparisons are likely to be (un)reliable. For transfer receipt, these measures may even allow researchers to get an impression of the likely direction of the bias, because we study over- and underreporting separately.

Survey users may also be able to use our findings to improve estimates. Our results can help them understand which analyses are likely to be biased and hence require corrections. For example, researchers interested in the relation between program receipt and income should be concerned that both stigma and recall issues are likely to increase with income. At the same time, salience of government programs is likely to decrease with income, all of which point toward more reporting errors at higher income levels. This relation makes it problematic to study the relation between

[25] This is possible when the model of interest is similar for the more reliable reporters or sample selection can be dealt with in other ways.

income and program receipt in survey data. Similar considerations can also help researchers understand whether and how they can correct for misreporting. Many corrections for misreporting rely on orthogonality conditions between errors and unit characteristics. By examining whether factors such as stigma, recall, salience or cooperation are likely to vary with their variables of interest, survey users can get an impression whether these assumptions are likely to hold. Doing so can help to gauge which corrections are (un)likely to improve estimates. For example, the considerations regarding the relationship of income to reporting errors above suggests that corrections for classical measurement error are unlikely to improve analyses of income and program receipt.

Our results can also help survey users make progress in the presence of non-classical measurement error, when few corrections are viable. Predictors of misreporting, such as our measure of cooperation, allow researchers to improve the accuracy of their estimates. For example, instrumental variable methods (e.g. Nguimkeu, Denteh and Tchernis 2019) require exclusion restrictions, i.e. variables that predict either misreporting or receipt, but not both. Our results above provide researchers with a menu of variables that predict misreporting, from which they can choose variables that are unlikely to predict their outcome of interest. See Denthe (2021) for an application.

Our results also have a number of implications for how survey producers can improve data accuracy. We show that response accuracy greatly varies between individuals and that this variation is partly predictable. Providing survey users with more information on the likely accuracy of a given response, such as the fraction of refused answers or the number of contact attempts, would allow them to deal with the problem of misreporting better, as we argue above. Investigating which additional paradata predict response errors and publishing such information would provide data users with a better means to address the problem. Thereby survey producers can still improve the data after the interviews have been conducted, i.e. after the damage from misreporting is done.

The theories we test also suggest ways to directly improve response accuracy by modifying survey design. For example, survey producers could try to exploit the salience effects we find to improve response accuracy. The results underline that the original intention of conducting the CPS around tax filing dates was good. In terms of government programs, the recall and salience results suggest that attempting to conduct surveys of program participation around their payment dates is likely to improve reports. Salience can also explain why smaller programs such as TANF are

reported less accurately, while large and salient programs such as social security are reported better. Taking these differences into account may provide survey producers with additional avenues to improve reporting. Our results also stress the importance of respondent cooperativeness (Bollinger and David, 2001). Understanding why people participate in surveys and what determines their cooperation may provide survey producers with means to improve survey accuracy, some of which are likely to be cheap compared to other measures. For example, US Census Bureau (2003) show that informing people that responding to the ACS is mandatory increases response rates from about 93 to 98 percent. Making other surveys mandatory is unlikely to be an option, but informing them of the importance and purpose of the survey may affect cooperation and thereby response accuracy.

However, survey producers mainly care more about data quality overall. Our results point to two important trade-offs that can help to inform survey design choices. First, our results on respondents' cooperation and misreporting suggest that efforts to increase response rates may worsen survey quality. Survey non-response has been the main topic of attention for decades and most efforts are dedicated to reducing non-response rates in surveys (Massey and Tourangeau 2013). However, households that are more reluctant to participate in surveys may also be households that, once they agree to participate, are more likely to misreport than others (Krosnick 1991, Groves and Couper 1998, Olson 2006, Tourangeau, Groves, and Redline 2010). To the extent that unit- and item-nonresponse are related, so that such behavior is proxied by our measures of cooperation, our results support the hypothesis of a trade-off between increasing response rates and improving response accuracy. Thus, the effect of simply increasing response rates on survey quality overall is ambiguous. Our finding that proxy interviews come at no cost in accuracy is important in this context, as it suggests that proxy interviews can reduce non-response without decreasing response accuracy.

Second, our results on recall also imply a trade-off between survey accuracy and detail. The steep increase of error rates with time since last receipt shows that there is scope to improve accuracy by adjusting reference periods, if feasible. For example, they suggest that the false positive rate could be reduced by more than 50 percent and false negatives reduced by 16 - 25 percent if the reference period of the CPS were the 12 months before the interview, as in the ACS. Survey error can be further reduced by using shorter reference periods. Surveys such as the CPS that are used for annual official statistics may have limited scope to adjust their reference periods,

but topical surveys such as FoodAPS or the surveys used in randomized control trials could substantially increase data accuracy by asking for current receipt. In general, our results suggest that more detailed questions, such as those on past events and their timing, yield more but less accurate information than simpler questions, such as those on current receipt. Survey producers should take this trade-off into account by considering how much detail they need to collect.

Awareness of such trade-offs is crucial to improve survey accuracy, particularly since most studies of survey design evaluate the effect of survey design features on specific outcomes (such as the response rate), without considering potential negative effects on other aspects of survey quality (such as response accuracy). Basic economic arguments imply that survey producers need to take such externalities into account in order to optimize survey quality overall. Understanding such trade-offs can also help survey users choose better among the available surveys, since they imply that using surveys with more detail or more emphasis on high response rates than required for their research may come at the expense of lower response accuracy.

The results above may also be useful for studies of other variables that have been shown to be affected by measurement error, such as income and education. While models of errors likely differ between variables, our findings still provide guidance on whether researchers should be concerned about misreporting and if so, what can still be learned from the contaminated data. For example, our results on recall and reference periods suggest that reports of whether or not an event happened are more reliable than the timing of the event. More generally, current and recent events are more likely to be accurately reported. To what extent our results are applicable to other variables remains an open question. Our evidence that whether survey participants respond accurately depends on the benefits and costs of providing a truthful answer may help researchers assess the applicability of our results to their case. Our findings clearly emphasize the importance of analyzing the causes of reporting error and demonstrate that linking survey data to more reliable measures of key variables is a feasible and promising avenue to do so.

VI. Conclusions

In this paper, we analyze reasons for measurement errors in the reports of government transfers. We link the ACS, the CPS, and the SIPP to administrative microdata on SNAP, TANF, and General Assistance from New York State to validate the survey responses. We study two types of errors in reports of program participation: false negative and false positive responses. While

there are many studies that explore causes of measurement errors in surveys, there are few that examine errors in several major surveys with extensive, high quality data. Past studies typically compare surveys to aggregate administrative data and thereby lack much of the detail found in microdata. Studies using microdata usually use one survey, study a single program, or use data that is 30 years old. We provide a more complete picture by comparing our results across different surveys and within surveys between different variables.

Our findings confirm several theories of cognitive factors leading to misreporting in surveys. We find that recall is an important source of response errors. Longer recall periods increase the probability that households fail to report program receipt. Problems of accurately recalling the timing of receipt, known as telescoping, are an important reason for overreporting. Our results also show that salience of the topic improves the quality of the answer. We provide evidence that respondents sometimes misreport when the true answer is likely to be known to them. Our results indicate that stigma indeed reduces reporting of program receipt. We show that cooperativeness affects the accuracy of responses in that interviewees who frequently nonrespond are more likely to misreport than other interviewees. In terms of survey design, we find no loss of accuracy from proxy interviews. Our results on survey mode effects are in line with the trade-off between non-response and accuracy that the previous literature points to.

Our results have implications for a broad area of research in economics using or producing surveys that are of importance beyond the case of government transfers and the specific surveys that we study. Our findings may allow data users to gauge the prevalence of errors in their data and select more reliable measures. That survey respondents tend to act like economic agents who consider the benefits and costs of their responses may help to assess data accuracy more broadly. The patterns we show can help to assess whether the errors are likely to be correlated with other variables of interest, thereby providing guidance on corrections for misreporting that are likely to reduce bias. We propose and test a measure of respondent cooperation at the unit level, which can be used to further assess and potentially reduce the impact of survey errors. Survey producers could make available other measures of the likely accuracy of a given response, such as the number of contact attempts. Data producers may also use our findings on reference periods and salience to improve response accuracy when deciding on the survey design. Finally, our results suggest important trade-offs data producers face. First, increasing response rates may not always be ideal, because additional efforts to elicit responses from reluctant households may bring inaccurate

respondents into the survey. Second, asking for more detailed information may yield less accurate responses. Taking such trade-offs into account can help survey producers optimize data accuracy in an economic way.

Our results and recommendations are broad enough to be applied in many settings where researchers believe that misreporting is a problem. For instance, similar issues of data quality have been found in health, crime, or earnings studies, to name a few. Researchers in these areas can be guided by our results to come up with methods to assess or address misreporting in their data. This study suggests that, if possible, linking survey data to administrative records, or other reliable sources of information is a promising way to reduce the problem of measurement errors in survey data.

References

Acs, Gregory, Katherin Ross Phillips, and Sandi Nelson. 2005. “The Road Not Taken?: Changes in Welfare Entry during the 1990s.” *Social Science Quarterly* 86: 1060-1079.

Almond, Douglas, Hilary W. Hoynes, and Diane Whitmore Schanzenbach. 2010. “Inside the War on Poverty: The Impact of Food Stamps on Birth Outcomes.” *Review of Economics and Statistics* 93 (2): 387–403.

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

Bachman, Jerald G., and Patrick M. O’Malley. 1981. “When Four Months Equal a Year: Inconsistencies in Student Reports of Drug Use.” *Public Opinion Quarterly* 45 (4): 536–48.

Belli, Robert F., Michael W. Traugott, and Mathew N. Beckman. 2001. “What Leads to Voting Overreports? Contrasts of Overreporters to Validated Voters and Admitted Nonvoters in the American National Election Studies.” *Journal of Official Statistics* 17 (4): 479–98.

Besley, Timothy, and Stephen Coate. 1992. “Understanding Welfare Stigma: Taxpayer Resentment and Statistical Discrimination.” *Journal of Public Economics* 48 (2): 165–83.

Biemer, Paul P., Robert M. Groves, Lars E. Lyberg, Nancy A. Mathiowetz, and Seymour Sudman, eds. 1991. "*Measurement Errors in Surveys*." 1 edition. New York: Wiley.

Black, Dan, Seth Sanders, and Lowell Taylor. 2003. “Measurement of Higher Education in the Census and Current Population Survey.” *Journal of the American Statistical Association* 98 (463).

Blank, Rebecca M. 2002. “Evaluating Welfare Reform in the United States.” *Journal of Economic Literature* 40 (4): 1105–66.

Blank, Rebecca M., and Patricia Ruggles. 1996. “When Do Women Use Aid to Families with Dependent Children and Food Stamps? The Dynamics of Eligibility Versus Participation.” *The Journal of Human Resources* 31 (1): 57–89.

Blattman, Christopher, Julian C. Jamison, and Margaret Sheridan. 2017. "Reducing crime and violence: Experimental evidence from cognitive behavioral therapy in Liberia." *American Economic Review* 107 (4): 1165-1206.

Bollinger, Christopher R., and Martin H David. 1997. “Modeling Discrete Choice with Response Error: Food Stamp Participation.” *Journal of the American Statistical Association* 92 (439): 827–35.

Bollinger, Christopher R. and Martin H. David. 2001. “Estimation with Response Error and Nonresponse: Food-Stamp Participation in the SIPP”, *Journal of Business and Economic Statistics*, 19:2, 129-141.

Bound, John, Charles Brown, and Nancy Mathiowetz. 2001. “Measurement Error in Survey Data.” In *Handbook of Econometrics*, edited by James J. Heckman and Edward Leamer, vol. 5:3705–3843. Amsterdam: Elsevier Science.

Brittingham, Angela, Roger Tourangeau, and Ward Kay. 1998. “Reports of Smoking in a National Survey: Data from Screening and Detailed Interviews, and from Self- and Interviewer-Administered Questions.” *Annals of Epidemiology* 8 (6): 393–401.

Bruckmeier, Kerstin, Gerrit Müller, and Regina T. Riphahn. 2014. “Who Misreports Welfare Receipt in Surveys?” *Applied Economics Letters* 21 (12): 812–16.

———. 2015. “Survey Misreporting of Welfare receipt—Respondent, Interviewer, and Interview Characteristics.” *Economics Letters* 129 (April): 103–7.

Butler, Joseph S., Richard Y. Burkhauser, Jean M. Mitchell, and Theodore P. Pincus. 1987. “Measurement Error in Self-Reported Health Variables.” *Review of Economics and Statistics* 69 (4): 644–50.

Cartwright, Ann, 1957. "The effect of obtaining information from different informants on a family morbidity inquiry." *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 6 (1):18-25.

Celhay, Pablo, Bruce D. Meyer and Nikolas Mittag. 2021a. “Errors in Reporting and Imputation of Government Benefits and Their Implications.” IZA DP 14396.

Celhay, Pablo, Bruce D. Meyer and Nikolas Mittag. 2021b. "Stigma in Welfare Programs". Unpublished Manuscript.

Chun, Asap Y., Michael D. Larsen, Gabriele Durrant, and Jerome P. Reiter, eds., 2021. "*Administrative Records for Survey Methodology*." New York: John Wiley & Sons.

Courtemanche, Charles, Augustine Denteh, and Rusty Tchernis. 2019. “Estimating the Associations between SNAP and Food Insecurity, Obesity, and Food Purchases with Imperfect Administrative Measures of Participation.*” Southern Economic Journal* 86 (1): 202–228.

Currie, Janet, Jeffrey Grogger, Gary Burtless, and Robert F. Schoeni. 2001. “Explaining Recent Declines in Food Stamp Program Participation [with Comments].” *Brookings-Wharton Papers on Urban Affairs*, January, 203–44.

Danielson, Caroline, and Jacob Alex Klerman. 2006. “Why Did the Food Stamp Caseload Decline (and Rise)?” http://www.rand.org/pubs/working_papers/WR386.html.

Da Silva, Damião N., Chris Skinner, and Jae Kwang Kim. 2016. “Using Binary Paradata to Correct for Measurement Error in Survey Data Analysis.” *Journal of the American Statistical Association* 111 (514): 526–37.

Davern, Michael, Bruce D. Meyer, and Nikolas Mittag. 2019. “Creating Improved Survey Data Products Using Linked Administrative-Survey Data.” *Journal of Survey Statistics and Methodology*. *7*(3): 440-463.

Davis, Darren W. 1997. “Nonrandom Measurement Error and Race of Interviewer Effects Among African Americans.” *Public Opinion Quarterly* 61 (1): 183–207.

Deaton, Angus. 1997. “The Analysis of Household Surveys : A Microeconometric Approach to Development Policy.” 17140. The World Bank.

de Leeuw, Edith D. 1992. "*Data Quality in Mail, Telephone and Face to Face Surveys*." Amsterdam: T. T. Publikaties.

DeMaio, Theresa J. 1984. “Social Desirability and Survey Measurement: A Review.” In *Surveying Subjective Phenomena*, C. F. Turner and E. Martin, 257–82. New York: Russell Sage Foundation.

Denteh, Augustine. 2021. "The effect of SNAP on obesity in the presence of endogenous misreporting." Unpublished Manuscript.

Fraker, Thomas, and Robert A. Moffitt. 1988. “The Effect of Food Stamps on Labor Supply: A Bivariate Selection Model.” *Journal of Public Economics* 35 (1): 25–56.

Fu, Haishan, Jacqueline E. Darroch, Stanley K. Henshaw, and Elizabeth Kolb. 1998. “Measuring the Extent of Abortion Underreporting in the 1995 National Survey of Family Growth.” *Family Planning Perspectives* 30 (3): 128–33, 138.

Gaskell, George D., Daniel B. Wright, and Colm A. O'Muircheartaigh. 2000. “Telescoping of Landmark Events: Implications for Survey Research.” *Public Opinion Quarterly* 64 (1): 77–89.

Gittleman, Maury. 2001. “Declining Caseloads: What Do the Dynamics of Welfare Participation Reveal?” *Industrial Relations: A Journal of Economy and Society* 40 (4): 537–70.

Gleason, Philip, Peter Schochet, and Robert Moffitt. 1998. “The Dynamics of Food Stamp Program Participation in the Early 1990s.” Mathematica Policy Research Reports ab95304cd2204323a950b50dd193aa7b. Mathematica Policy Research. https://ideas.repec.org/p/mpr/mprres/ab95304cd2204323a950b50dd193aa7b.html.

Gray, Percy G. 1955. “The Memory Factor in Social Surveys.” *Journal of the American Statistical Association* 50 (270): 344–63.

Grogger, Jeffrey. 2002. “The Behavioral Effects of Welfare Time Limits.” *The American Economic Review* 92 (2): 385–89.

Groves, Robert M. 2006. “Nonresponse Rates and Nonresponse Bias in Household Surveys.” *Public Opinion Quarterly* 70 (5): 646–75.

———. 2011. “Three Eras of Survey Research.” *Public Opinion Quarterly* 75 (5): 861–71.

Groves, Robert M., and Mick P. Couper. 1998. “*Nonresponse in Household Interview Surveys.*” New York: John Wiley and Sons.

Groves, Robert M., Floyd J. Fowler, Mick P. Couper, James M. Lepkowski, Eleanor Singer, and Roger Tourangeau. 2009. “*Survey Methodology.*” Wiley Series in Survey Methods. New York: John Wiley and Sons.

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

Gundersen, Craig, and Victor Oliveira. 2001. “The Food Stamp Program and Food Insufficiency.” *American Journal of Agricultural Economics* 83 (4): 875–87.

Hoynes, Hilary W., and Diane W. Schanzenbach. 2012. “Work Incentives and the Food Stamp Program.” *Journal of Public Economics* 96 (1–2): 151–62.

Johnson, Timothy, and Michael Fendrich. 2005. “Modeling Sources of Self-Report Bias in a Survey of Drug Use Epidemiology.” *Annals of Epidemiology* 15 (5): 381–89.

Kaminska, Olena, Allan L. McCutcheon, and Jaak Billiet. 2010. “Satisficing Among Reluctant Respondents in a Cross-National Context.” *Public Opinion Quarterly* 74 (5): 956–84.

Kanuk, Leslie, and Conrad Berenson. 1975. “Mail Surveys and Response Rates: A Literature Review.” *Journal of Marketing Research* 12 (4): 440–53.

Karlan, Dean, and Jonathan Zinman. 2008. “Lying About Borrowing.” *Journal of the European Economic Association* 6 (2-3): 510–21.

Karlan, Dean S. and Jonathan Zinman., 2012. "List randomization for sensitive behavior: An application for measuring use of loan proceeds." *Journal of Development Economics*, *98*(1): 71-75.

Krosnick, Jon A. 1991. “Response Strategies for Coping with the Cognitive Demands of Attitude Measures in Surveys.” *Applied Cognitive Psychology* 5 (3): 213–36.

Lyberg, Lars, and Daniel Kasprzyk. 1991. “Data Collection Methods and Measurement Error: An Overview.” In *Measurement Errors in Surveys*. Hoboken, NJ: John Wiley & Sons, Inc.

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

Massey, Douglas S., and Roger Tourangeau. 2013. “Where Do We Go from Here? Nonresponse and Social Measurement.” *The ANNALS of the American Academy of Political and Social Science* 645 (1): 222–36.

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

Meyer, Bruce D. and Nikolas Mittag. 2019b. “Misreporting of Government Transfers: How Important are Survey Design and Geography?” *Southern Economic Journal*. 86(1): 230-253.

Meyer, Bruce D. and Nikolas Mittag. 2021a. “An Empirical Total Survey Error Decomposition Using Data Combination.” *Journal of Econometrics*. 224(2): 286-305.

Meyer, Bruce D. and Nikolas Mittag 2021b. "Combining Administrative and Survey Data to Improve Income Measurement." In *Administrative Records for Survey Methodology*, edited by Asap Y. Chun, Michael D. Larsen, Gabriela Durrant and Jerome P. Reiter, cp. 12: 297-322. New York: John Wiley and Sons.

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., Nikolas Mittag and Robert M. Goerge. 2021. “Errors in Survey Reporting and Imputation and Their Effects on Estimates of Food Stamp Program Participation.” *Journal of Human Resources*.

Meyer, Bruce D., and Derek Wu. 2018. “The Poverty Reduction of Social Security and Means-Tested Transfers.” *ILR Review* 71 (5): 1106–53.

Mittag, Nikolas. 2019. “Correcting for Misreporting of Government Benefits.” *American Economic Journal: Economic Policy* 11(2): 142-16.

Moffitt, Robert A. 1983. “An Economic Model of Welfare Stigma.” *American Economic Review* 73 (5): 1023–35.

Moffitt, Robert A. 2014. “Economics of Means-Tested Transfer Programs in the United States, Volume 1.” *NBER*, December. http://papers.nber.org/books/moff14-1.

Moore, Jeffrey C. 1998. “Self/Proxy Response Status and Survey Response Quality, A Review of the Literature.” *Journal of Official Statistics* 4 (2): 155–72.

Neter, John, and Joseph Waksberg. 1964. “A Study of Response Errors in Expenditures Data from Household Interviews.” *Journal of the American Statistical Association* 59 (305): 18–55.

Niehaus, Paul and Sandip Sukhtankar. 2013. “Corruption dynamics: The golden goose effect.” *American Economic Journal: Economic Policy*, 5 (4): 230-269.

Nguimkeu, Pierre, Augustine Denteh, and Rusty Tchernis. 2019. "On the estimation of treatment effects with endogenous misreporting." *Journal of Econometrics* 208(2): 487-506.

Olson, Kristen. 2006. “Survey Participation, Nonresponse Bias, Measurement Error Bias, and Total Bias.” *Public Opinion Quarterly* 70 (5): 737–58.

Ribar, David C. 2005. “Transitions from Welfare and the Employment Prospects of Low-Skill Workers.” *Southern Economic Journal* 71 (3): 514–33.

Ridder, Geert, and Robert A. Moffitt. 2007. “Chapter 75 The Econometrics of Data Combination.” In *Handbook of Econometrics*, edited by James J. Heckman and Edward E. Leamer, Vol. 6, Part B:5469–5547. Amsterdam: Elsevier Science.

Sharp, Jeff S., and Lazarus Adua. 2010. “Examining Survey Participation and Response Quality: The Significance of Topic Salience and Incentives.” *Survey Methodology* 36 (1): 95–109.

Sirken, Monroe. 1999. *Cognition and Survey Research*. Vol. 322. Wiley-Interscience.

Sudman, Seymour, and Norman M. Bradburn. 1973. “Effects of Time and Memory Factors on Response in Surveys.” *Journal of the American Statistical Association* 68 (344): 805–15. doi:10.1080/01621459.1973.10481428.

———. 1974. *Response Effects in Surveys: A Review and Synthesis*. Aldine Publishing Company.
Tamborini, Christopher R, and ChangHwan Kim. 2013. “Are Proxy Interviews Associated with Biased Earnings Reports? Marital Status and Gender Effects of Proxy.” *Social Science Research* 42 (2): 499–512.
Tourangeau, Roger. 1984. “Cognitive Science and Survey Methods.” *Cognitive Aspects of Survey Methodology: Building a Bridge between Disciplines*, 73–100.
Tourangeau, Roger, Robert M. Groves, and Cleo D. Redline. 2010. “Sensitive Topics and Reluctant Respondents: Demonstrating a Link between Nonresponse Bias and Measurement Error.” *Public Opinion Quarterly* 74 (3): 413–32. doi:10.1093/poq/nfq004.
Tourangeau, Roger, and Kenneth A Rasinski. 1988. “Cognitive Processes Underlying Context Effects in Attitude Measurement.” *Psychological Bulletin* 103 (3): 299.
Tourangeau, Roger, Lance J. Rips, and Kenneth A Rasinski. 2000. "*The Psychology of Survey Response*." Cambridge: Cambridge University Press.
Tourangeau, Roger, and Ting Yan. 2007. “Sensitive Questions in Surveys.” *Psychological Bulletin* 133 (5): 859–83. doi:10.1037/0033-2909.133.5.859.
US Census Bureau. 2003. “Report 3: Testing the Use of Voluntary Methods.” Meeting 21st Century Demographic Data Needs— Implementing the American Community Survey. http://www.census.gov/library/working-papers/2003/acs/2003_Griffin_01.html.
———. 2006. “Design and Methodology: Current Population Survey.” 66. U.S. Census Bureau.
———. 2008. “Survey of Income and Program Participation: User’s Guide.” 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.
Zabel, Jeffrey E., 1998. "An analysis of attrition in the Panel Study of Income Dynamics and the Survey of Income and Program Participation with an application to a model of labor market behavior." *Journal of Human Resources* 33(2): 479-506.

Tables

TABLE 1: PROGRAM PARTICIPATION QUESTIONS IN THE CPS, ACS, AND SIPP

Survey Program ***Years***	**Text of Question**
	The household respondent, any adult above 15 years old in the household, answers the following questions:
CPS SNAP *(2007/2010)*	Did (you/ anyone in this household) get food stamps or a food stamp benefit card at any time during 200x? *If the answer is "No" then the survey asks:* At any time during 200x, even for one month, did (you/ anyone in this household) receive any food assistance from (State Program name)?
CPS SNAP *(2011)*	At any time during 2010, even for one month, did (you/ anyone in this household) receive any food assistance from (State Program name) or a food assistance benefit card (such as State EBT card name)?
CPS SNAP *(2012 onwards)*	Did (you/ anyone in this household) get food stamps or a food stamp benefit card at any time during 20xx? *If the answer is "No" then the survey asks:* At any time during 20xx, even for one month, did (you/ anyone in this household) receive any food assistance from (State Program name) or a food assistance benefit card (such as State EBT card name)?
CPS PA *(2007/2013)*	At any time during 200x, even for one month, did (you/ anyone in this household) receive any CASH assistance from a state or county welfare program such as (State Program Name)? *The survey insists with an additional question:* Just to be sure, in 200x, did anyone receive CASH assistance from a state or county welfare program, on behalf of children in the household?
	The household respondent, any adult above 15 years old, answers the following question:
ACS SNAP *(2007/2009)*	In the past 12 months, did anyone in this household receive Food Stamps or a Food Stamp benefit card?
ACS	

Survey Program ***Years***	**Text of Question**
SNAP *(2010/2012)*	*In years 2010 to 2012 the survey adds the following interviewer instructions:* "Include government benefits from the Supplemental Nutrition Assistance Program (SNAP). Do NOT include WIC or the National School Lunch Program."
ACS SNAP *(2013 onwards)*	*In year 2013 the ACS changed the question to:* In the past 12 months, did you or any member of this household receive benefits from the Food Stamp Program or SNAP (the Supplemental Nutrition Assistance Program)? *The survey adds the following interviewer instructions:* "Do NOT include WIC, the School Lunch Program, or assistance from food banks."
	For every household member, the survey collects the following information from the household respondent:
ACS PA *(2007/2013)*	Any public assistance or welfare payments from the state or local welfare office in the past 12 months *The survey asks for the amount directly without a previous Yes/No question about receipt.*
	Every member of the household who is 15 years old or older answers each of the following questions:
SIPP SNAP *(2004,2008)*	*In each wave the survey asks first for the first reference month:* Since [reference month 1] 1st, was [household member] authorized to receive food stamps? *For each of the following reference months the survey asks:* Earlier I recorded that [household member] received Food Stamps in [reference month 1]. Did [he or she] also receive Food Stamps in [reference month 2], [reference month 3], [reference month 4], or [reference month 5]?
SIPP PA *(2004,2008)*	Since [reference month 1] 1st, Did [household member] receive any CASH assistance from a state or county welfare program, such as [State Program name for TANF] or AFDC? *If the answer is "Yes" then the survey asks:* Which program was that? (What do you call it?) *Then the survey moves to the following question for all "Yes" and "No" respondents:* How about General Assistance or General since [reference month 1]? Did [household member] receive any short-term cash assistance since [reference month 1] 1st to tide [him or her] over when [he or she] needed it to help [him or her] stay off welfare, or for an emergency?

Survey Program *Years*	Text of Question
	Then the survey asks a similar question for the reference month 1 for the following programs: "short-term cash assistance", "child support", "transportation assistance", "child care assistance", "food assistance", "clothing assistance", "housing assistance", "other state welfare program". *For each of the following reference months the survey asks:* Earlier I recorded that [household member] received [State program] in [reference month 1]. Did [he or she] also receive those payments in [reference month 2], [reference month 3], [reference month 4], or [reference month 5]?

Notes: For the CPS, State Programs listed for Public Assistance include cash payments from: welfare or welfare-to-work programs, Temporary Assistance for Needy Families program, Aid to Families with Dependent Children, General Assistance/Emergency Assistance program, Diversion Payments, Refugee Cash and Medical Assistance program, General Assistance from the Bureau of Indian Affairs, or Tribal Administered General Assistance. For more information about questions in the CPS visit http://www.nber.org/cps/cpsmar11.pdf. For more information about specific questions in the ACS visit https://www.census.gov/programs-surveys/acs/methodology/questionnaire-archive.html. For more information about specific questions and fieldwork manual for the SIPP visit https://www.census.gov/programs-surveys/sipp/tech-documentation/questionnaires.html. Source: CPS: US Census Bureau (2006), ACS: US Census Bureau (2014), and SIPP: US Census Bureau (2008).

TABLE 2: SNAP PROGRAM ERROR RATES AND DESCRIPTIVE STATISTICS FOR DETERMINANTS OF ERRORS

		CPS			**ACS**			**SIPP**		
	Sample	*Mean*	*SD*	*N*	*Mean*	*SD*	*N*	*Mean*	*SD*	*N*
False negative rate	True recipients	0.373	0.484	2,993	0.249	0.432	79,831	0.180	0.384	4,330
False positive rate	True non-recipients	0.012	0.110	12,735	0.011	0.104	457,061	0.013	0.112	18,726
Months since last receipt	True recipients	1.526	1.769	2,993	1.558	1.888	79,831	1.088	0.419	4,330
Receipt before reference period	True non-recipients	0.007	0.082	12,735	0.029	0.168	368,508	0.009	0.096	17,354
Receipt after reference period	True non-recipients	0.027	0.163	12,735						
Months of receipt	True recipients	10.164	3.198	2,993	10.092	3.233	79,831	3.632	0.840	4,330
Monthly amount received ($100)	True recipients	2.924	1.893	2,993	3.058	2.026	79,831	2.845	1.898	4,330
Participation rate in ZIP Code	True recipients	0.506	0.312	2,472	0.483	0.312	79,627	0.339	0.203	4,290
Low cooperation with survey	All respondents	0.202	0.401	15,728	0.154	0.361	537,432	0.198	0.399	23,056
Very low cooperation with survey	All respondents	0.030	0.171	15,728	0.090	0.286	537,432	0.075	0.264	23,056
Proxy interview	All respondents							0.140	0.347	23,056
Future attrition	All respondents							0.476	0.499	23,056
CATI Interview	All respondents	0.131	0.337	15,728	0.078	0.268	537,432			
CAPI interview	All respondents				0.382	0.486	537,432	1	0	23,056

Notes: This table reports descriptive statistics for all variables used in the Probit models. All samples are restricted to households that respond to the question whether they receive SNAP. True non-recipients are our sample for the false positive analysis, which includes all households that do not receive SNAP benefits according to our linked administrative measure (but our Probit models below exclude observations for which the participation rate in the ZIP code is missing). True recipients are our sample for the false negative analysis, which includes all households that receive SNAP according to our linked administrative measure (but our Probit models below exclude households for which information on receipt before the reference period is missing, because they were interviewed early in the time period covered by our administrative records). Low cooperation with survey is an indicator for being between the 75th and 90th percentile of the distribution of the number of questions the respondent refused to answer (see Appendix Table 1). Very low cooperation indicates being above the 90th percentile of this distribution. Note that the percentiles refer to the distribution in the entire sample that includes imputed observations, so the fraction in our sample differs from the 15 and 10 percent the percentiles would suggest. Observations are weighted using survey weights adjusted for PIK probability using inverse probability weighting. The Census Bureau has reviewed this data product for unauthorized disclosure of confidential information and has approved the disclosure avoidance practices applied to this release, authorization date: October 6, 2017.

TABLE 3: PUBLIC ASSISTANCE ERROR RATES AND DESCRIPTIVE STATISTICS FOR DETERMINANTS OF ERRORS

		CPS			ACS			SIPP		
	Sample	*Mean*	*SD*	*N*	*Mean*	*SD*	*N*	*Mean*	*SD*	*N*
False negative rate	True recipients	0.590	0.492	779	0.587	0.492	14,610	0.458	0.499	840
False positive rate	True non-recipients	0.003	0.059	14,939	0.013	0.115	493,329	0.004	0.066	22,229
Months since last receipt	True recipients	2.533	2.926	779	2.308	2.939	14,610	1.196	0.615	840
Receipt before reference period	True non-recipients	0.005	0.070	14,939	0.024	0.154	401,086	0.004	0.061	20,634
Receipt after reference period	True non-recipients	0.009	0.092	14,939						
Months of receipt	True recipients	8.226	3.886	779	8.247	3.874	14,610	3.435	0.997	840
Monthly amount received ($100)	True recipients	5.463	3.682	779	5.567	3.844	14,610	5.618	3.605	840
Participation rate in ZIP Code	True recipients	0.141	0.094	659	0.134	0.094	14,570	0.090	0.062	831
Low cooperation with survey	All respondents	0.202	0.402	15,718	0.153	0.360	507,939	0.198	0.398	23,069
Very low cooperation with survey	All respondents	0.030	0.170	15,718	0.065	0.246	507,939	0.075	0.264	23,069
Proxy interview	All respondents							0.140	0.347	23,069
Future attrition in the SIPP	All respondents							0.476	0.499	23,069
CATI Interview	All respondents	0.133	0.339	15,718	0.080	0.271	507,939			
CAPI interview	All respondents				0.389	0.487	507,939	1	0	23,056

Notes: This table reports descriptive statistics for all variables used in the Probit models. All samples are restricted to households that respond to the question whether they receive PA. True non-recipients are our sample for the false positive analysis, which includes all households that do not receive PA benefits according to our linked administrative measure (but our Probit models below exclude observations for which the participation rate in the ZIP code is missing). True recipients are our sample for the false negative analysis, which includes all households that receive PA according to our linked administrative measure (but our Probit models below exclude households for which information on receipt before the reference period is missing, because they were interviewed early in the time period covered by our administrative records). Low cooperation with survey is an indicator for being between the 75^{th} and 90^{th} percentile of the distribution of the number of questions the respondent refused to answer (see Appendix Table 1). Very low cooperation indicates being above the 90^{th} percentile of this distribution. Note that the percentiles refer to the distribution in the entire sample that includes imputed observations, so the fraction in our sample differs from the 15 and 10 percent the percentiles would suggest. Observations are weighted using survey weights adjusted for PIK probability using inverse probability weighting. The Census Bureau has reviewed this data product for unauthorized disclosure of confidential information and has approved the disclosure avoidance practices applied to this release, authorization date: October 6, 2017.

TABLE 4: PROBIT ESTIMATES OF THE DETERMINANTS OF FALSE NEGATIVE REPORTS BY PROGRAM AND SURVEY
(AVERAGE MARGINAL EFFECTS)

	Food Stamp Program			**Public Assistance**		
	CPS	*ACS*	*SIPP*	*CPS*	*ACS*	*SIPP*
Months since last receipt	0.0362***	0.0325***	0.0219	0.0357***	0.0303***	0.0912***
	(0.0071)	(0.0010)	(0.0135)	(0.0075)	(0.0020)	(0.0322)
Months of receipt	-0.0149***	-0.0190***	-0.0498***	-0.0125**	-0.0139***	-0.0284
	(0.0032)	(0.0005)	(0.0072)	(0.0055)	(0.0013)	(0.0223)
Monthly amount received ($100)	-0.0068	-0.0030***	-0.0115**	-0.0110**	-0.0019	0.0150**
	(0.0070)	(0.0011)	(0.0057)	(0.0043)	(0.0012)	(0.0067)
Participation rate in ZIP Code	-0.0765**	-0.0066	0.0939	-0.3190*	-0.1293**	0.4899
	(0.0327)	(0.0068)	(0.0604)	(0.1901)	(0.0526)	(0.4684)
Low cooperation with survey	0.0556***	0.0100**	0.0063	-0.0236	0.0249**	0.0052
	(0.0201)	(0.0046)	(0.0272)	(0.0389)	(0.0121)	(0.0656)
Very low cooperation with survey	0.3054***	0.0381***	-0.0084	-0.0198	0.0385**	-0.0116
	(0.0783)	(0.0052)	(0.0296)	(0.1479)	(0.0162)	(0.1001)
Proxy interview			0.0430			-0.0144
			(0.0271)			(0.0743)
Future attrition			0.0084			0.0316
			(0.0225)			(0.0617)
CATI interview		0.0813***			0.0659***	
		(0.0052)			(0.0152)	
CAPI interview	0.1534***	0.1644***		0.0328	0.1427***	
	(0.0345)	(0.0039)		(0.0612)	(0.0094)	
Observations	2,472	79,627	4,290	659	14,570	831

Notes: This table reports average marginal effects on the probability of a false negative report in each survey and program. All regressions are based on the sample of non-imputed observations and control for family type, number of adults, number of children, household head's sex, age, education, race, disability condition, and citizenship status, whether households are rural, whether household head speaks English poorly, reported receipt of other programs, and year. See note to tables 2 and 3 for further information on definitions. Observations are weighted using survey weights adjusted for PIK probability using inverse probability weighting. Standard errors in parentheses (clustered at the household level in the SIPP). *** p<0.01, ** p<0.05, * p<0.10. The Census Bureau has reviewed this data product for unauthorized disclosure of confidential information and has approved the disclosure avoidance practices applied to this release, authorization dates: October 26, 2016, and October 6, 2017.

TABLE 5: PROBIT ESTIMATES OF THE DETERMINANTS OF FALSE POSITIVE REPORTS BY PROGRAM AND SURVEY
(AVERAGE MARGINAL EFFECTS)

	Food Stamp Program			**Public Assistance**		
	CPS	*ACS*	*SIPP*	*CPS*	*ACS*	*SIPP*
Receipt before reference period	0.0089 (0.0061)	0.0072*** (0.0009)	0.0199*** (0.0051)	0.0012 (0.0018)	0.0115*** (0.0010)	0.0111*** (0.0026)
Receipt after reference period	0.0061 (0.0045)			0.0056*** (0.0020)		
Low cooperation with survey	0.0045** (0.0022)	0.0011* (0.0006)	0.0047 (0.0035)	0.0035*** (0.0010)	-0.0005 (0.0006)	0.0037** (0.0016)
Very low cooperation with survey	0.0046 (0.0061)	0.0032*** (0.0007)	0.0081 (0.0056)	0.0022 (0.0027)	0.0045*** (0.0007)	-0.0001 (0.0015)
Proxy interview			0.0038 (0.0036)			-0.0010 (0.0014)
Future attrition			0.0005 (0.0029)			0.0017 (0.0012)
CATI interview		0.0017** (0.0007)			-0.0098*** (0.0008)	
CAPI interview	0.0055* (0.0032)	0.0045*** (0.0006)		0.0039** (0.0019)	-0.0095*** (0.0005)	
Observations	12,735	368,508	16,174	14,817	401,086	19,446

Notes: This table reports average marginal effects on the probability of a false positive report in each survey and program. All regressions are based on the sample of non-imputed observations and control for family type, number of adults, number of children, household head's sex, age, education, race, disability condition, and citizenship status, whether households are rural, whether household head speaks English poorly, reported receipt of other programs, and year. See note to tables 2 and 3 for further information on definitions. Observations are weighted using survey weights adjusted for PIK probability using inverse probability weighting. Standard errors in parentheses (clustered at the household level in the SIPP). *** p<0.01, ** p<0.05, * p<0.10. The Census Bureau has reviewed this data product for unauthorized disclosure of confidential information and has approved the disclosure avoidance practices applied to this release, authorization dates: October 26, 2016, and October 6, 2017.

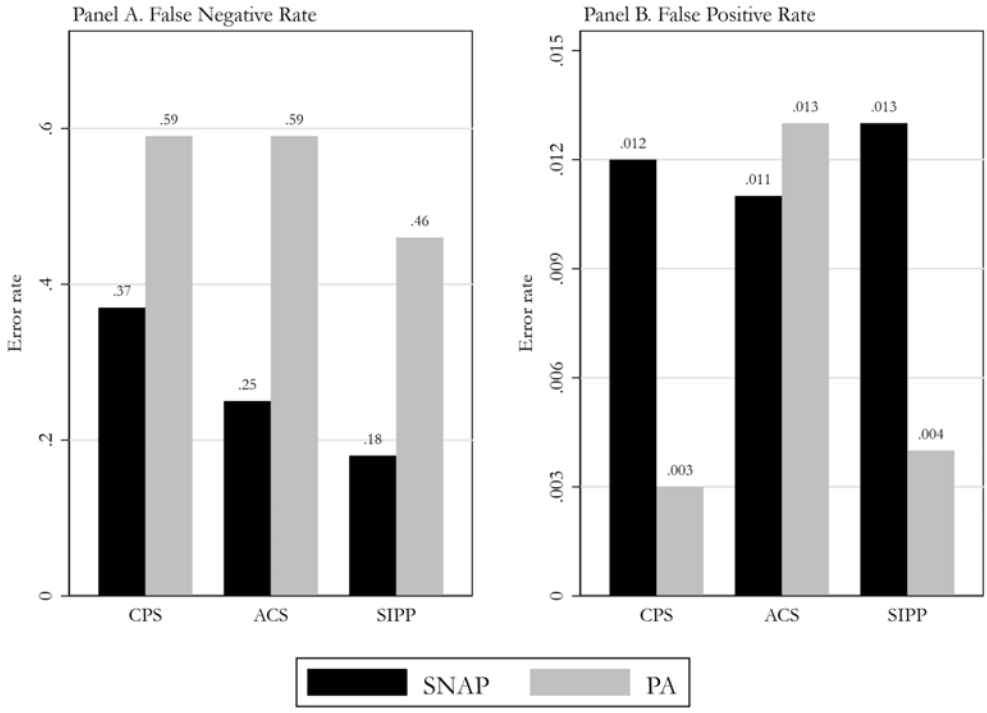
TABLE 6: PROBIT ESTIMATES OF THE DETERMINANTS OF ERRORS BY INTERVIEW MODE IN THE AMERICAN COMMUNITY SURVEY (AVERAGE MARGINAL EFFECTS)

	False Negative Response				**False Positive Response**			
	Food Stamp Program		*Public Assistance*		*Food Stamp Program*		*Public Assistance*	
	Non Mail	Mail	Non Mail	Mail	Non Mail	Mail	Non Mail	Mail
Months since last receipt	0.0387***	0.0222***	0.0319***	0.0289***				
	(0.0017)	(0.0009)	(0.0027)	(0.0027)				
Receipt before reference period					0.0076***	0.0073***	0.0080***	0.0156***
					(0.0018)	(0.0007)	(0.0013)	(0.0013)
Months of receipt	-0.0193***	-0.0172***	-0.0147***	-0.0115***				
	(0.0008)	(0.0005)	(0.0016)	(0.0020)				
Monthly amount received ($100)	-0.0029*	-0.0031**	-0.0027*	0.0008				
	(0.0016)	(0.0012)	(0.0014)	(0.0019)				
Participation rate in ZIP Code	-0.0221**	0.0214***	-0.1419**	-0.0310				
	(0.0096)	(0.0061)	(0.0636)	(0.0810)				
Low cooperation with survey	0.0115*	0.0001	0.0133	0.0252	0.0028	0.0024***	-0.0023**	0.0016**
	(0.0069)	(0.0045)	(0.0159)	(0.0168)	(0.0020)	(0.0005)	(0.0011)	(0.0007)
Very low cooperation with survey	0.0654***	0.0110**	0.0089	0.0494***	0.0016	0.0010**	0.0084***	0.0047***
	(0.0112)	(0.0043)	(0.0342)	(0.0183)	(0.0014)	(0.0005)	(0.0018)	(0.0008)
Observations	37,270	42,357	8,299	6,271	99,147	269,361	119,985	281,101
Error rate	0.298	0.156	0.618	0.497	0.017	0.007	0.012	0.014

Notes: This table reports average marginal effects on the probability of a reporting error in the ACS for SNAP and PA. All regressions are based on the sample of non-imputed observations and control for family type, number of adults, number of children, household head's sex, age, education, race, disability condition, and citizenship status, whether households are rural, whether household head speaks English poorly, reported receipt of other programs, and year. See note to tables 2 and 3 for further information on definitions. Observations are weighted using survey weights adjusted for PIK probability using inverse probability weighting. Standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.10. The Census Bureau has reviewed this data product for unauthorized disclosure of confidential information and has approved the disclosure avoidance practices applied to this release, authorization date: October 6, 2017.

Figures

FIGURE 1: ERROR RATES IN THE REPORTED RECEIPT OF PA AND SNAP IN THE CPS, ACS, AND THE SIPP.



Notes: Error rates are for the non-imputed samples of each survey and program and use survey weights adjusted for PIK probability using inverse probability weighting. See Tables 2 and 3 for the underlying numbers and observation counts. The Census Bureau has reviewed this data product for unauthorized disclosure of confidential information and has approved the disclosure avoidance practices applied to this release, authorization dates: October 26, 2016, and October 6, 2017.

APPENDIX TABLE 1: VARIABLES USED TO CONSTRUCT A MEASURE OF COOPERATIVENESS USING IMPUTATION INDICATORS OF OTHER VARIABLES IN EACH SURVEY

Current Population Survey		American Community Survey		SIPP	
Variable name	**Variable label**	**Variable name**	**Variable label**	**Variable name**	**Variable label**
nxtres	Main reason for moving	fbld	Units in structure	pubhse	Public housing
caid	Medicaid coverage	fele	Elctricty cost	gvtrnt	Government program
care	Medicare coverage	hfl	Heating fuel type	egyast	Energy assistance
discs	Retiremente for a reason	mvy	How long living here (years)	bmnth	Month of birth
dishp	Health problem or dissability	refr	Refrigerator	byear	Year of birth
dis_yn	Disability income other than Social Security …	rms	Number of rooms	educate	Education
div_yn	Dividends received	rwat	Hot and cold running water	race	Race
ed_yn	Educational assistance	stov	Stove or range	ms	Marital Status
ern_yn	Earnings from longest job	tel	Telephone in unit	citizen	Citizenship
fin_yn	Financial assistance	ten	Housing tenure	receipt	Program receipt
hea	Health status self reported	toil	Flush toilet		
hi	Covered by employer or health plan	veh	Vehicles		
int_yn	Interest payments	wat	Water cost		
othstper	Covered by other type of health insurance	ybl	Year structure first built		
oth	Covered by any other type of health insurance	age	Age		
out	Covered by the health plan of someone …	anc	Ancestry recode		
priv	Covered by a private plan purchased directly	cit	US Citizenship		
retyn	Pension or retirement income other than …	ddrs	Difficulty dressing		
rnt_yn	Rent income received	db	Ambulatory difficulty		
seyn	Self employment income	dear	Hearing difficulty		
ssi_yn	Supplemental Security income received	deye	Vision or hearing difficulty		
ss_yn	Social Security payments received	dout	Difficulty going out		
suryn	Survivor's benefits other than Social Sec …	dphy	Physical difficulty		
ucyn	Unemployment compensation benefits received	drem	Difficulty remembering		
vet_yn	Veterans payments received	esr	Raw labor-force status		
wc_yn	Worker's compensation payments received	hins	Health insurance (binary)		
workyn	Worked at job or business during year	his	Hispanic, Detailed		
wsyn	Any wage and salary reported	mar	Marital status		
hengast	Energy assistance benefits	mig	Mobility status		
paw_yn	Public assistance received	frac	Race		
hfoods	Food stamps recipients	rel	Relationship		
hloren	Reduced rent, Federal, State, or local …	ss	Social security		
hpubli	Public housing project	pa	Public assistance		
hunit	Number of units in this structure	ret	Retirement		
axage	Age				
amrital	Marital Status				
hga	Educational attaintment				
lfrs	Labor force status				
lflj	Last work for pay at a regular job or …				

To construct our measure of cooperation with the survey, we first compute the fraction of the questions listed above that the respondent refused to answer. We then create indicators whether a respondent was above the 90th percentile of the distribution of the fraction of questions refused to answer (very low cooperation with the survey) and another indicator whether the respondent was between the 75th and the 90th percentile of this distribution (low cooperation with the survey). The quantile cutoffs are based on the sample that includes imputed observations. This choice ensures that the cooperation measure does not change between the SNAP and PA models, which differ in their sample due to differences in item non-response.