THE UNIVERSITY OF CHICAGO

ESSAYS IN PUBLIC ECONOMICS

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE IRVING B. HARRIS
GRADUATE SCHOOL OF PUBLIC POLICY STUDIES
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

BY
PABLO CELHAY BALMACEDA

CHICAGO, ILLINOIS
JUNE 2016

Para Manuela y Salvador...

las razones de todo

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

This dissertation is dedicated to study government programs from different perspectives. Chapter 1 presents evidence of measurement error in the report of program participation in Food Stamps and Public Assistance in different surveys of the U.S. Measurement error is often the largest source of bias in survey data, yet little is known about the determinants of such errors, making 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 different causes of survey error using high quality validation data from three major surveys in the U.S. that are linked to administrative data on government transfers. The differences between survey and administrative records show that up to six out of ten cash welfare recipients are missed by surveys. We find that survey design and post-processing as well as misreporting by respondents affect survey errors systematically. Imputation for missing data induces substantial error. Our results on respondent behavior confirm several theories of misreporting, e.g. that errors are related to salience of receipt, respondent's degree of cooperation, forward and backward telescoping, event recall, and the stigma of reporting participation in social programs. Our results provide guidance on the conditions under which survey data are likely to be accurate and suggest different ways to control for survey errors. Chapter 2 investigates whether fixed costs of adjustment as opposed to low returns explain why better quality care practices diffuse slowly in the medical industry. Using a randomized field experiment, the results show that temporary financial incentives paid to health clinics for the early initiation of prenatal care nudged providers to test and develop new data driven strategies to locate and encourage likely pregnant women to seek care in the first trimester of pregnancy. These innovations raised the rate of early initiation of prenatal care by 34% while the incentives were being paid in the treatment period. Following health clinics over time the findings illustrate that this increase persisted for at least 24 months after the incentives ended. In the absence of incentives, it is in the clinics' interest to provide better prenatal care but

learning and experimenting with new methods is too costly. The temporary incentives help to overcome initial costs and increase productivity in the long run. Despite the large increases in early initiation of prenatal care, there are no effects on health outcomes. Chapter 3 explores the effects of a conditional cash transfer program in Bolivia that pays mothers between $7 and $18 per health visit for prenatal checkups, skilled birth attendance and preventive health care checkups for children up to 24 months. Using municipal-level data from national health systems, the results show that the geographic variation in the penetration of the program coincided with a 12% reduction in the rate of stillbirths. Different tests assess that this relation is likely to be causal. This result is supported by a quasi-experimental analysis using data from a nationally representative household survey that shows that program beneficiaries experienced higher rates of early detection of pregnancies, total number of prenatal care visits, and a higher rate of skilled birth attendance and postpartum car. For children 0 to 2 years old, the program increased the number of checkups and reduced the prevalence of anemia, though we find no evidence of longer-term impacts on stunting or wasting. Since transfer amounts represent a small proportion of household consumption, we posit that health impacts are generated mainly through increased utilization of preventive healthcare rather than an income effect. The intervention is highly cost-effective, at $716.1 per DALY averted, equivalent to 29% of GDP per-capita.

ACKNOWLEDGMENTS

First and foremost my infinite gratitude goes to my wife, Manuela, who has put up with five unimaginable winter seasons in Chicago, with a graduate student inside the house. I also received throughout these five years a great support from my family in Chile. This dissertation is also dedicated to them.

This dissertation would not be the same without the help of many people at the Harris School. I had the fortune to work for almost three years with a great mentor, Bruce Meyer. He has put up with my unattractive writing style, polishing it a long the way and always giving great advice to achieve the "James Bond Movie" paper we are always after. I have learned a lot from him these years, and for sure have a lot to learn yet. I am also indebted to Dan Black and his CPE workshop, birthplace, and burying ground, of many ideas for research that graduate students had. Many thanks to Bob Lalonde and Jeff Grogger for comments, meetings, and attention to my work. A million thanks to Cynthia Cook-Conley who has been a great support in these five years.

The work presented here would not have been possible without the help of my co-authors: Paula Giovagnoli, Sebastian Martinez, Nikolas Mittag, and Christel Vermeersch. I am particularly grateful to Paul Gertler from UC Berkeley, who has been a major influence throughout the last years in my thinking about economics, data analysis, and academia in general.

Finally, I would like to thank staff at the Census RDC, the Plan Nacer team in Argentina, and everybody involved in the evaluation of the Bono Juana Azurduy in UDAPE, Government of Bolivia. All of them made possible to work with the data used throughout this dissertation.

CHAPTER 1

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

In collaboration with Bruce D. Meyer and Nikolas Mittag[*†]

1.1 Introduction

For decades, household surveys have been one of the most important tools for empirical work in economics and other social sciences. Nationally representative surveys in the U.S. are the main source of official statistics such as the unemployment, poverty, and health coverage rates. Likewise, a growing fraction of studies are producing their own data using household surveys. Whether one uses data from secondary sources or produces it, results ultimately rely on how well surveys measure the topic of interest.

The quality of survey data, however, has been declining steadily in the recent years.[1] Households are more reluctant to participate in surveys, and if they agree to participate they are more likely to refuse to answer particular questions or give inaccurate responses. Non-response rates have been increasing for nearly all surveys in the U.S. and a large literature

∗. Meyer: Harris School of Public Policy Studies, University of Chicago, 1155 E. 60th Street, Chicago, IL 60637, (email: bdmeyer [at] uchicago.edu). Mittag: CERGE-EI, joint workplace of Charles University Prague and the Economics Institute of the Academy of Sciences of the Czech Republic, Politickýchv veznu 7, Praha, Czech Republic, (email: nikolas.mittag [at] cerge-ei.cz).

†. The data in this paper was provided by the U.S. Census Bureau. All analyses were performed at the Research Data Center of the Census by researchers with Special Sworn Status. All results and opinions are those of the authors.

1. Survey data quality has been falling steadily in recent years (see Massey and Tourangeau, 2013 and Meyer et al., 2015). The quality of observational data has been a topic of interest for many years and has inspired an extensive literature in statistics and the social sciences. There are many reviews available. Some examples are Biemer et al. (2011), Bound et al. (2001), Alwin (2007) , and Groves et al. (2011).

has attempted to understand the causes and consequences of the trend.[2] Less is known about measurement error, i.e. the difference between the recorded response of a household that participates in a survey and a measure of truth for the same variable.[3]

In this paper we study measurement error in surveys and analyze different theories about its nature. In particular, we study measurement error in the report of participation in government programs. While a few past studies have focused on the topic, they have usually compare surveys to aggregate administrative data, and have therefore been unable to test the many hypotheses that require micro-data. Studies using micro-data generally use one survey, study a single program, or use data that is 30 years old.[4] We provide a more powerful examination of the reasons for errors by comparing results across different surveys and different programs. We also suggest ways to account for measurement error in surveys, which should be relevant to a broad area of research in economics that uses surveys.

Response errors in survey data are common. They can be found in variables that are unlikely to elicit uneasiness, such as education (Black et al., 2003), as well as potentially illegal or stigmatizing variables, such as questions related to drug use by teenagers (Johnson and Fendrich, 2005) or self-reported health status (Butler et al., 1987). Moreover, Celhay et al. (2016) find that measurement error generates large biases in survey data, with the magnitude of the bias being more than three times that of survey non-response in our situation (see Figure 1.1). Furthermore, others have found that response errors are not

2. See Groves (2006), Groves (2011) and Massey and Tourangeau (2013) for a review of the unit non-response literature. Survey non-response bias reflects systematic differences between people who agree to participate in the survey and people who decline to participate in the survey. In 2010 the Russell Sage Foundation commissioned the National Research Council's Committee on National Statistics to assemble a panel of experts dedicated to study the causes and consequences of increasing non-response rates. Furthermore, there is evidence showing that non-response rates in household surveys in the U.S. have been steadily increasing in the past decade (see Meyer et al., 2015).

3. We refer to measurement error as the difference between the recorded data and a measure of truth for the same variable.

4. See David (1962), Marquis and Moore (1990), Bollinger and David (1997), Meyer et al. (2015), and Meyer et al. (2015). See Bound et al. (2001) for a review and Bruckmeier et al. (2014) for a study of misreporting of welfare receipt in Germany.

independent of other characteristics of respondents, which will generally bias both causal and descriptive estimates, when either the dependent or independent variable is measured with error since they are binary.[5] One reason for the dearth of empirical research on reasons for measurement errors in surveys is that reliable measures of "truth" for survey variables are rare. We work with high quality administrative records of the Food Stamp (SNAP) and Public 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 ASEC supplement 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).

We study two types of errors in binary variables: false negative responses (failures of true recipients to report or errors of omission) and false positive responses (reported receipt by those who are not in the administrative data or errors of commission).[6] Figure 1.2 summarizes the aggregate error rates in our data. For example, the probability of a false negative response is 62.9% for participation in cash welfare, i.e. more than six out of ten cash welfare recipients are recorded as non-recipients in the CPS. In addition, 0.61% of households that are not true recipients are recorded as having received welfare aid.[7] Comparing across surveys and programs provides a richer analysis of the determinants of errors. Since each survey is a random sample of the same population we should expect similar error rates. However, Figure 1.2 shows important differences across surveys suggesting that survey design can substantially affect error rates. In fact, the three surveys we use differ in many dimensions.[8]

5. See Bollinger and David (1997), Meyer et al. (2015), and Bound et al. (2001).

6. False negatives refer to true recipients of a program who do not report receiving aid when asked. False positives refer to observations that are not in the administrative data but report receiving aid in the survey.

7. The false positive rate may seem negligible in relative terms; however, the absolute sample count (123) is approximately a quarter of the total sample that reports receipt in the survey (456).

8. For instance, the main purpose of the CPS is to collect information on labor force statistics, while the SIPP invests highly on gathering accurate information about participation in social programs. Likewise, the ACS is a short-duration survey that respondents are legally obligated to respond, while the other two are

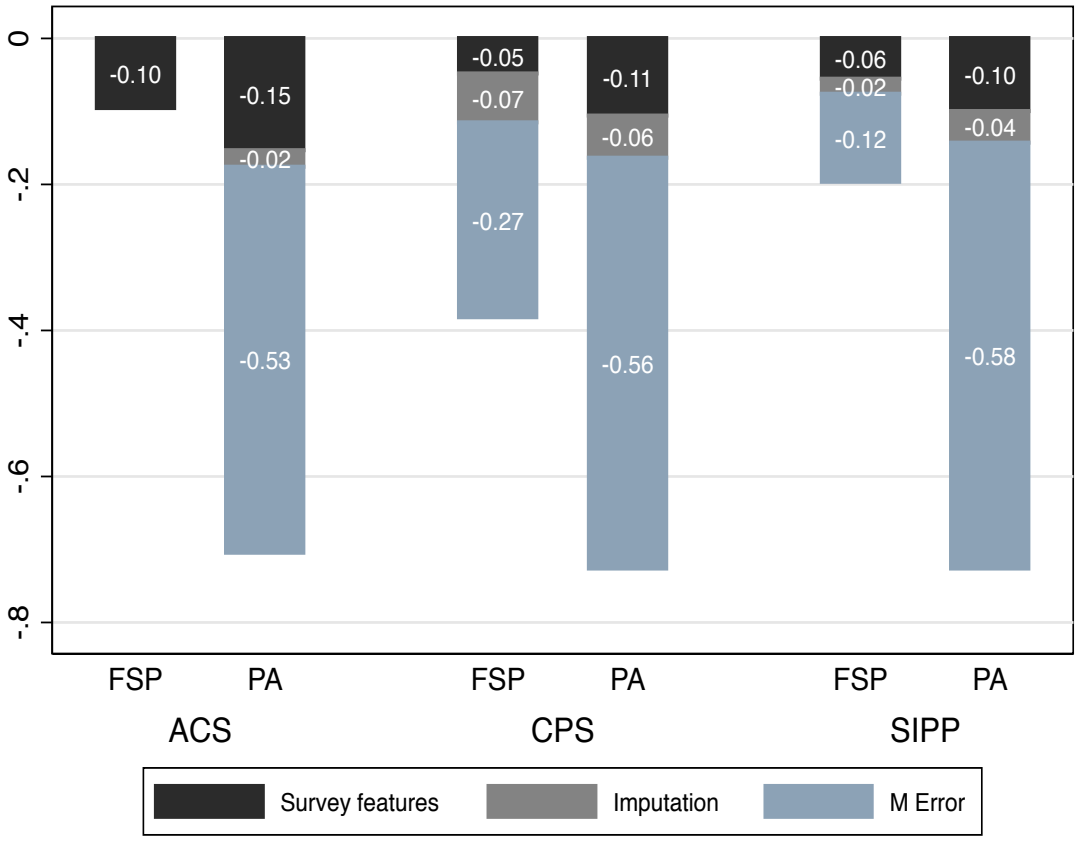

Figure 1.1: Decomposition of Proportional Bias in Dollars Received into Its Sources Using Microdata

Source: Authors' elaboration based on New York State data for 2007-2012 from Celhay et al. (2015).

Notes: We calculate the bias due to the combination of errors in coverage, weighting, and unit nonresponse as the ratio of weighted administrative program dollars received by all linked households in the Current Population Survey to total administrative dollars paid out minus one. We calculate the bias due to item nonresponse as weighted dollars imputed to those not responding to the benefit question minus the dollars actually received by these households as a share of total dollars paid out. Finally, we calculate the bias due to measurement error as the dollars recorded by non-imputed respondents minus true dollars received as a share of total dollars paid out. The surveys are the American Community Survey (ACS), the Current Population Survey (CPS), and the Survey of Income and Program Participation (SIPP). Food stamp dollars received are not reported in these years of the ACS.

What leads to survey errors? The literature often divides different causes into those related

voluntary. In addition, the SIPP is a longitudinal survey where respondents are visited every four months for four consecutive years, which adds a higher burden to interviewees when compared to the other two surveys.

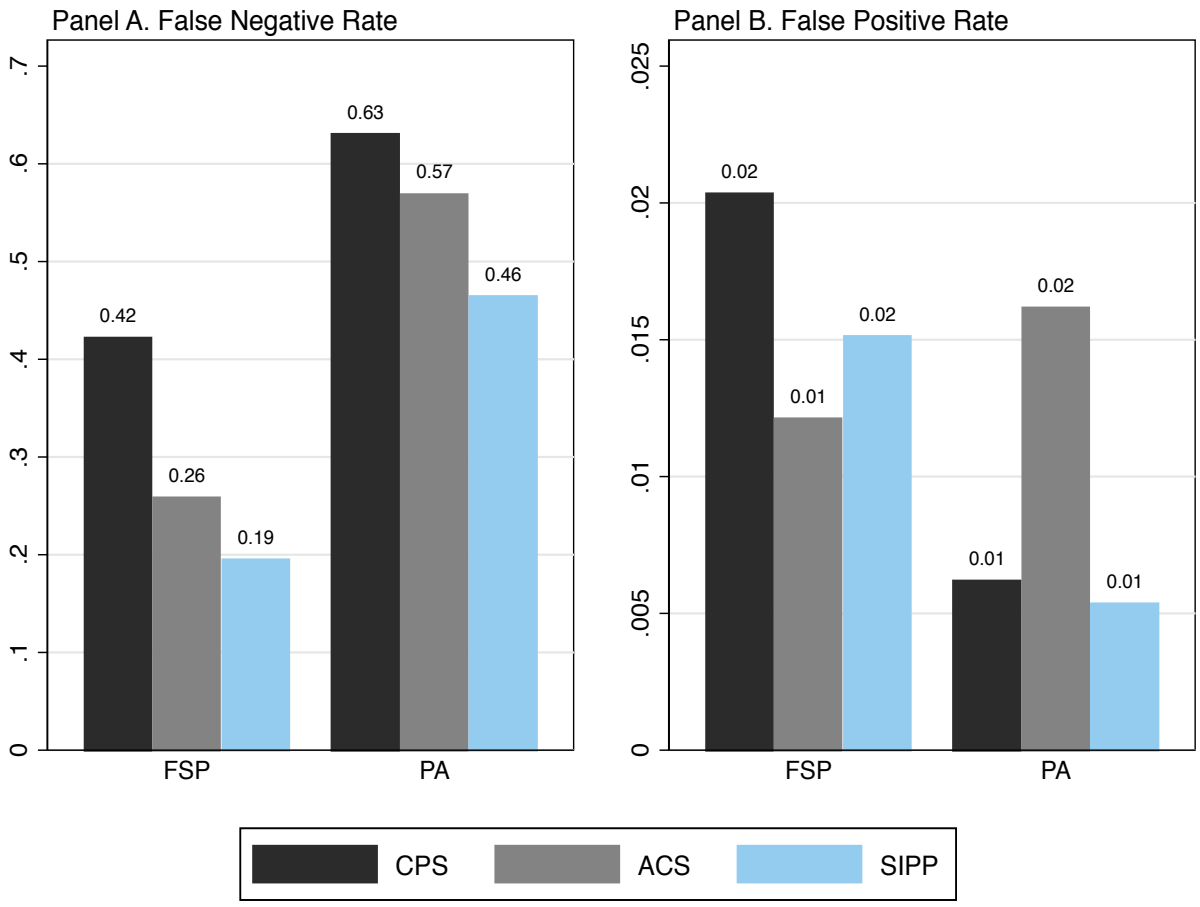

Figure 1.2: Error Rates in the Reported Receipt of Food Stamps and Public Assistance in the CPS, ACS, and the SIPP

Source: Authors' elaboration based on New York State data for 2007-2012 from Celhay et al. (2015).

Notes: Panel A shows the false negative rate in the response of participation for Food Stamps (FSP) and Public Assistance (PA) for the Current Population Survey (CPS), the American Community Survey (ACS) and the Survey of Income and Program Participation (SIPP). Panel B shows the same numbers for the false positive rate in the response of program participation.

to survey design features and those related to specific characteristics of respondents (Sudman and Bradburn, 1974; Groves et al., 2009). For instance, imputation of missing data may lead to misclassification of individual observations, households may be reluctant to reveal private information to an interviewer, they may have trouble accurately recalling events from the

past or may simply dislike surveys and misreport answers to get through them faster.[9]

We test how different variables that resemble theories of misreporting relate to each of the two errors. Our findings have several implications. We show that errors in recorded receipt are much higher among observations with imputed receipt. This result suggests that researchers may want to use the subsample of non-imputed households and use own methods to control for potentially non-random item non-response.[10] Our results also show that harder to reach households and those interviewed in person have higher error rates. We also find that households that are less cooperative with other sections of the survey are more likely to give incorrect answers. As such, measurement error could be modeled including variables that measure number of contact attempts and thus proxies for overall survey cooperation as in Bollinger and David (2001). Importantly, our results suggest that increasing response rates may not be always ideal since additional efforts to convince otherwise reluctant households may screen bad respondents into the survey.[11] Our results also show that reference periods in questionnaires create recall errors and telescoping effects. In addition, households that are more dependent on government transfers are better reporters on average, suggesting that salient events are less subject to response errors.[12] Finally, we find that households that live in ZIP codes with higher program participation rates are more likely to reveal true participation implying that people that live in areas with lower stigma are less likely to underreport "socially undesirable behavior".[13]

9. An extensive discussion of reasons for response error in surveys can be found in Sudman and Bradburn (1974). Mathiowetz et al. (2001) provide a survey on reasons for measurement error in studies of the low-income population. Other studies include Meyer et al. (2015) and Celhay et al. (2016) who analyze the effects of imputation on measurement error; Gaskell et al. (2000) who look at recall and telescoping effects in surveys, and Krosnick (1991) who analyzes survey errors as consequences of respondents' to get through a survey faster.

10. This is in line with other researchers that have previously studied the effect of imputation on earnings (see for example Lillard et al., 1986 and Hirsch and Schumacher, 2004).

11. See Groves and Couper (2012), Olson (2006), and Tourangeau et al. (2010) for a discussion on how efforts to increase survey response rates can bring less cooperative respondents into the survey.

12. See Mathiowetz et al. (2001).

13. See Sudman and Bradburn (1974), DeMaio (1984) and Tourangeau and Yan (2007) for studies on

This paper is related to a large literature that uses large-scale household surveys to study the effects and determinants of government programs, or research that uses social statistics to study the income distribution or the poverty rate.[14] A common approach in these studies is to ignore measurement error or assume that measurement error is uncorrelated with the true value of the mismeasured variable and uncorrelated with other covariates in the model.[15] Our results show that these assumptions are rejected by survey data implying that measurement error may have important consequences for estimation (see Figure 1.3).[16] Our findings are also related to a large literature that analyzes data quality in household surveys.[17] In particular, we use high quality validation data, different surveys, and multiple programs, providing extensive evidence of what causes measurement error. Understanding what leads to errors in surveys should be relevant to both users of survey data and to a large number of researchers in economics who gather their own surveys, particularly in the areas of experimental and development economics. Our results are broad enough to be applied to many examples where researchers believe that measurement error from misreporting is a problem, e.g. health, crime, or earnings studies. Researchers in these areas can be guided by our results to come up with creative methods to control for misreporting in their own data.

This paper is structured as follows. In section 1.2 we describe the data and how we link administrative records to survey data. In section 1.3 we define and describe the errors in reporting program participation across surveys and programs. In section 1.4 we summarize

sensitive topics and misreporting.

14. Some examples of this literature are Blank and Ruggles (1996), Deaton (1997), Fraker and Moffitt (1988), Gleason et al. (1998), Currie et al. (2001), Gundersen and Oliveira (2001), Blank (2002), Danielson and Klerman (2006), Gittleman (2001), Grogger (2002), Hoynes and Schanzenbach (2012), Almond et al. (2011), and Moffitt (2016).

15. See Griliches et al. (1986), Hausman et al. (1998), and Katz and Katz (2010).

16. Bound et al. (2001) and Hausman (2001) provide an extensive survey on the topic. Some examples on consequences of measurement error are Duncan and Hill (1985), Bollinger (1996), Hyslop and Imbens (2001), Black et al. (2000), and Celhay et al. (2016).

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

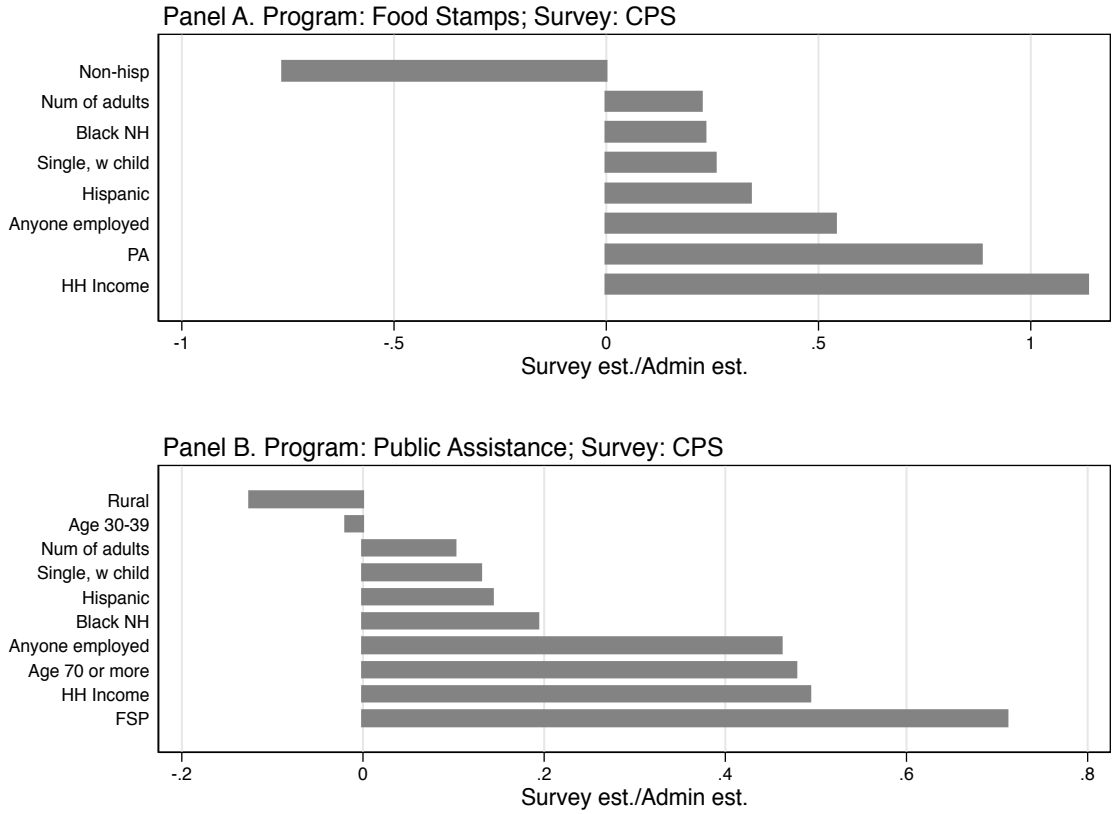

Figure 1.3: Ratio of Marginal Effects Estimated using Survey Receipt to Marginal Effects Estimated using True Receipt. Programs: Food Stamps and Public Assistance. Survey: CPS.

Source: Authors' elaboration based on New York State data for 2007-2012 from Celhay et al. (2015).

Notes: These figures shows the ratio of marginal effects obtained from a model of program participation where survey report is used as the dependent variable (numerator) to the same model estimated using true receipt as the dependent variable (denominator).

different theories that explain measurement error and in section 1.5 we test these theories. In section 1.6 we discuss how our results can help researchers account for measurement errors. Section 1.7 concludes.

1.2 Data

1.2.1 Survey Data

We study measurement error of program receipt in the American Community Survey (ACS), the Annual Social and Economic Supplement of the Current Population Survey (CPS), and the Survey of Income and Program Participation (SIPP). Each survey contains basic demographic information, as well as receipt of government assistance, labor force participation, and education, among others. In this section, we provide a short review of each survey. The specific questions that we work with can be found in 1.1.

Table 1.1: Program receipt questions in each survey

Survey	**Program**	**Text in questionnaire**	**Reference period**
CPS	Food Stamps	Did anyone in this household get ... at any time in 20XX?	12 months in the past calendar year
	Public Assistance	Did household member receive ... in 20XX?	
ACS	Food Stamps	IN THE PAST 12 MONTHS, did you or any member of this household receive benefits from ...?	Last 12 months as of the month of the interview
	Public Assistance	Any public assistance or welfare payments from the state or local welfare office in the past 12 months	
SIPP	Food Stamps	In month X did HOUSEHOLD MEMBER receive any income from ...?	Each of the last 4 months as of the month of the interview
	Public Assistance	In month X did HOUSEHOLD MEMBER receive any income from ...	

The ACS surveys approximately 2.5 percent of the U.S. population each year, with the interviews spread across all months of the year. It is the largest household survey in the U.S., with more than 290,000 households selected each month to participate. The ACS questionnaire is similar to the long-form decennial censuses that it replaced and is administered by mail, telephone, or face-to-face interview. In terms of information on government transfers,

it asks for participation in the FSP but not for amount received, but asks for both receipt and amount received for PA. For both programs, the questions in the ACS refer to the last 12 months prior to the interview date. We use the ACS for the years 2008 through 2012.

The CPS is one of the most important surveys in the U.S. It is the official source of labor force statistics in the country with more than 60,000 households participating in the survey each month of the year. We use the ASEC supplement of the CPS, which also is the official source of income information used to calculate the annual poverty rate 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 in the FSP and PA programs during the previous calendar year, 2007-2012 in our case.

Finally, the SIPP is the highest quality source of information on poor households and the effects of government income transfers. We employ the 2004 and 2008 panels of the SIPP that consist of approximately 50,000 households who are followed over 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 the SNAP and PA. In our analysis, we aggregate program participation and total amounts received within a four month wave for each household and analyze each wave as a separate cross section. We use the SIPP 2004 (waves 10 through 12) and the SIPP 2008 (all waves).

1.2.2 Administrative Data and Data Linkage

We link the three surveys to administrative records on FSP 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 the Food Stamp Program (FSP), Temporary Assistance for Needy Families (TANF), and General Assistance in New York

State from January 2007 through December 2012.[18,19] Each record in the data corresponds to monthly payments transferred 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. The records are from actual payments, and appear to be accurate. For the FSP, for example, the overall total dollars from our administrative records differs from official aggregate outlays by less than a percent in all years. These data have been previously used by Meyer et al. (2015) and Meyer and Mittag (2015), who further discuss its 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.[20] 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) [21] and attached to the corresponding records in our data. A PIK is obtained for over 99 percent of the administrative records from each program. Our unit of analysis is a household.[22] We consider a household to have a PIK

18. 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 exceed those of TANF both nationally and in New York. Likewise, the FSP has experienced a large increase in its caseload in recent years, making it one of the largest in-kind transfer programs of the country.

19. The data contain information separately for different components of PA, which we aggregate into one in our analysis.

20. See Wagner and Layne (2014) for an explanation on how linking between administrative data and surveys at the US Census Bureau is possible. See Ridder and Moffitt (2007) for a review on linking administrative records to surveys.

21. There is a one-to-one correspondence between PIKs and social security numbers; PIKs are used to protect the anonymity of individuals in the data. All analyses in this paper were done at the Chicago Census Research Data Center by researchers with Census Special Sworn Status.

22. The household is a natural unit of analysis for FSP since eligibility and program receipt are determined at the household level. However, PA receipt is determined individually so that the aggregation at the household level underestimates the rate of program participation. This issue has been also discussed in David and Bollinger (2000) and Meyer et al. (2015).

if a PIK was obtained for anyone in the household. The PIK rates at the household level are 93% in the ACS, 91% in the CPS, and 95% in the SIPP. We use the household as the unit of analysis, which is logical given the sharing of resources among members, but using households also insures a high rate of data linkage.[23] Since the administrative data have records for each recipient person, we are able to link the information from a program case to the household if anyone in the household who is recorded as receiving program benefits has a PIK. In order to account for the incomplete linking, we multiply the household weights by the inverse of the predicted probability of any household member having a PIK (see e.g. Wooldridge, 2007). The coefficients of the Probit model we use to predict these probabilities in each of the surveys are available upon request. As the high rate of PIK-linking suggests, our results do not appreciably change when using the adjusted household weights.

The breadth of these data is unusual in studies that rely on validation data to assess measurement error in household surveys. A common feature of studies that use administrative microdata linked to surveys is that they are typically available for a short time period, for only one survey, one program, and/or a small subsample of respondents. As such, any conclusion from these studies may be particular to the sample used and period of analysis, compromising any generalization of their results.

With our data, we are able to study measurement error at the household level using six years of data, calendar years 2007 through 2012, and compare results across two programs and three surveys using more than 90% of the sample in each of them. After linking the ACS, CPS, and the SIPP to the universe of cases that receive aid from the FSP or PA in New York State, our final sample size is 543,538 households in the ACS, 18,064 households in the CPS, and 24,997 household-wave observations in the SIPP. These are large samples in

23. A survey household may contain more than one individual receiving transfers from FSP or PA and include individuals who are not receiving any transfer from these programs. Since not all individuals in a household are necessarily assigned a PIK, there is a bias in aggregate error rates. As it is further discussed in Meyer et al. (2015) we are able to "... accurately determine what share of true recipient survey households report receipt, but we cannot determine what share of true recipient assistance units report receipt".

comparison to other studies of the topic.[24] In addition, more than 99% of the administrative records have a PIK, so that any bias from imperfect linking comes from survey observations for which a PIK is not available.[25]

Moreover, working with the universe of cases that receive any aid from the FSP or PA allows us to have an accurate measure of truth, unlike other variables such as income obtained from IRS records where the measure of true income is less clear. Likewise, the longitudinal form of the administrative records allows us to know whether a household received aid from the FSP 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.

We are also able to compare error rates across the two programs within the same survey using the same linkage method, which controls for any sample composition bias from differential linkage, while also holding constant any error from survey design. Finally, using calendar years 2007 through 2012 gives us a large sample to work with, while it also allows examining measurement error in a more recent period than other studies and at a time when the FSP caseload was growing rapidly.

1.3 Misclassification errors: definition

Our main approach is to compare survey response of program participation, to a measure of true program participation. The three surveys ask about program participation at any point in their reference period, i.e. a measure y_{ij}, that equals one if household i reported receipt from program j during the reference period. Our linked administrative data allow us to match the definition of program participation in survey data when calculating a measure of true participation, y^*_{ij}, that equals one if household i received any transfer from program j during the reference period. Using these two measures, we define two classification errors:

24. See Marquis and Moore (1990), Bollinger and David (1997), and Meyer et al. (2015).

25. As we discussed above, we adjust our results by the probability of having a PIK using IPW methods. Our results hold without this adjustment.

false negatives ($y_{ij} = 0|y^*_{ij} = 1$), and false positives ($y_{ij} = 1|y^*_{ij} = 0$). The false negatives correspond to households that are true recipients of a program, but fail to report receipt when asked in the survey. The false positives are households that did not receive benefits from program j during the reference period of the survey, but were recorded as having received aid from that program.

In the next section we discuss different hypotheses that explain the magnitude of the error rates as well as why they differ across programs and surveys, and in section V we test these ideas empirically.

1.4 Reasons for Misclassification of Program Participation Variables

In this section we shortly review different theories of misreporting in surveys. In order to organize the discussion, we divide the theories of misreporting into components of survey design and data post-processing (survey features) that affect survey errors, and theories related to respondent's characteristics.[26][27]

1.4.1 Survey features

Questions of survey design that have been shown to affect the extent of survey errors arise at all stages of the survey process. For instance, at the very first stage, incomplete sampling frames can bias survey estimates (Kish, 1965), and even when working with complete frames, selective unit non-response could increase errors from survey data (Groves, 2004). Post processing problems can also affect survey errors trough data editing, data entry, and the design of weights. The analysis of how different components of survey design affect survey

26. For a more comprehensive review of the evidence on the relation between survey design and measurement error see Alwin (2007) and Groves et al. (2009).

27. For a review on respondent behavior and response quality see Sudman and Bradburn (1974), Sirken (1999) and Bound et al. (2001), to name a few.

data quality has been analyzed under the "Total Survey Error" literature. Groves and Lyberg (2010) provide an excellent review. We focus on a particular set of survey components for which we can present evidence of their association with measurement error: mode of interview, interviewer effects, proxy interviews, and imputation of missing data.

The first issue on which we provide evidence below is the choice of survey mode, i.e. whether the data is collected in face-to-face interviews, interviews assisted by computer software (CAPI), telephone interviews (CATI), or by self-administered mail-back questionnaire. The choice of interview mode usually involves factors such as costs, coverage error, and questionnaire length (Groves et al., 2009, pp. 149). However, one of the main concerns discussed in the literature involves the trade-off between increasing participation rates in the survey and improving response quality of participants (Kanuk and Berenson, 1975; Lyberg and Kasprzyk, 1991). In a meta-analysis of the effects of interview mode on survey quality, de Leeuw (1992) finds that while non-mail interviews have better response rates, the effects of interview mode on response quality can be large, but their magnitude and direction depends on the context. For instance, questions that address sensitive behavior are reported better in self administered surveys as opposed to telephone interviews (Tourangeau and Smith, 1996). This has been found in the report of drug use (Aquilino and Sciuto, 1990), sexual behavior (Turner et al., 1998), and other sensitive behavior (Newman et al., 2002).

The modes of data collection clearly differ in how much they restrict interviewer impact. In mail surveys the interviewer is absent and cannot play a role -either positive or negative- in the question-answer process. In telephone interviews, which have a limited channel capacity, interviewers have potentially less impact on respondent behavior than in face-to-face interviews. In any case, both in-person and telephone interviews rely on the performance of interviewers (Bradburn, 2016). Empirical studies on interviewer effects show large heterogeneity and many mechanisms through which interviewers can affect survey quality.[28] For

28. Some examples are O'Muircheartaigh and Campanelli (1998), Pickery et al. (2001), Essig and Winter (2009), and Kalwij (2010).

instance, they can worsen response in sensitive questions (Tourangeau and Yan, 2007) or increase response rates and response accuracy if their characteristics are similar to that of the respondents' (Bruckmeier et al., 2015).

In addition to the choice of interviewers, an important survey design question is who to interview. A costly option is to interview each household member. A less costly alternative is to interview only one person per household, asking this reference person the survey questions for every member of the household. Such proxy interviews, in which the respondent provides information about other individuals, are also commonly used to reduce non-response (Todorov and Kirchner, 2000). The evidence on the effect of proxy interviews on survey accuracy is mixed. On one hand, they can improve on self-reports for sensitive questions (Tourangeau et al., 2000). On the other hand, proxy interviews may increase error rates, because proxy respondents are less informed about person-specific issues (Tamborini and Kim, 2013).[29] Some have even argued that proxy interviews do not differ substantially from self-reports (Moore, 1998) in terms of error rates, but it is not clear whether this is evidence that proxy interviews do not matter or that the effects the other studies examine offset each other.

Issues of survey design that affect survey accuracy continue to arise after the data has been collected, since survey post-processing affects the degree of error. In most surveys, the largest effect of survey post-processing is the treatment of missing data. This is commonly done by imputation, i.e. by assigning values to non-respondents through methods such as the hot deck.[30] Although imputation can be useful to obtain unbiased aggregate means from survey data there is ample evidence showing that imputation can induce substantial error in survey data (Meyer et al., 2015; Celhay et al., 2016) and in estimates derived from them

29. See also Cartwright (1957) and Blair et al. (2004), for other examples on the effects of proxy interviewing.

30. For a description of how surveys impute missing data see US Census Bureau (2006) for the CPS, US Census Bureau (2008) for the SIPP, and US Census Bureau (2014) for the ACS.

(Lillard et al., 1986; Bollinger and Hirsch, 2006).

1.4.2 Respondents' and Event-specific Characteristics

Respondent behavior and their characteristics can also lead to survey errors. It is common in this literature to conceptualize survey response as depending on four steps: a) comprehension of the question, b) retrieval of information, c) judgement, and d) communication of final response.[31] From an economic perspective, a respondent should provide a correct answer if revealing the truth yields higher utility, i.e. benefit net of cost, than an incorrect or no answer. A respondent has to be willing to incur the cost of the effort invested to comprehend the question and to attempt to recall the correct information or look it up somewhere. Even if the answer is known, so that recall costs are low, the respondent also has to be willing to reveal her answer to a stranger, increasing the costs of reporting from social stigma if the topic of the survey is sensitive (Karlan and Zinman, 2008).

Comprehension of the question and the ability to retrieve information are related to cognitive behavior of the respondent (Tourangeau, 1984). Interpretation or comprehension issues are mostly dealt with through questionnaire design and training of interviewers (Mathiowetz et al., 2001). However, there is less leverage to address other issues of respondent behavior in surveys. For instance, the amount of effort invested to retrieve accurate information about a question may be lower if the event is more salient, i.e. more frequent or highly present in a respondent's mind. As such, respondents with a higher knowledge about the topic in question may give better answers (Sharp and Adua, 2010) and be more willing to participate in the survey (Groves et al., 2004). In addition, respondents may have to exert more effort to retrieve information in questions that use reference periods. Such questions assume that respondents can situate in different time-specific intervals of time during the survey (Groves et al., 2009, pp. 231). However, respondents will make mistakes in remembering time-specific

31. For a detailed discussion of this model and related literature see Tourangeau (1984), Tourangeau and Rasinski (1988), Tourangeau et al. (2000), and Groves et al. (2009, pp. 217).

events. They may bring forward past events into the reference periods, known as telescoping (Gaskell et al., 2000), or may forget an experience that occurred long ago from the time of the interview (Gray, 1955; Eisenhower et al., 2004).

On the other hand, judgment and final communication of an answer affect survey error because they make accurate answers costly to respondents due to their willingness to answer questions even when the correct answer is known to them (Krosnick, 1991). A frequently discussed cost of reporting true behavior is stigma, i.e. the cost of providing socially undesirable answers. Stigma has been shown to cause respondent to provide inaccurate answers to sensitive question, even in cases where truth is known to the respondent (DeMaio, 1984; Tourangeau et al., 2010). Previous research has found that there is important misreporting in surveys on topics such as abortion (Fu et al., 1998), drug use (Brittingham et al., 1998) and sexual orientation (Coffman et al., 2013). While the literature has emphasized stigma as a reason for failures to provide undesirable answers, social desirability can also lead to overreporting of desirable answers such as voting behavior (Belli et al., 2001).

Table 1.2: Descriptive Statistics for the Food Stamps (SNAP) Program Sample

		CPS			ACS			SIPP		
	Sample	Mean	SD	N	Mean	SD	N	Mean	SD	N
False positive rate	True non-recipients	0.02	0.141	14,525	0.012	0.109	461,756	0.015	0.122	20,226
False negative rate	True recipients	0.421	0.494	3,539	0.257	0.437	81,772	0.194	0.395	4,771
Survey response imputed	All sample	0.13	0.336	18,064	0.011	0.103	543,528	0.072	0.258	24,997
CATI Interview	Non-imputed sample	0.131	0.337	15,728	0.078	0.268	537,432			
CAPI Interview	Non-imputed sample				0.382	0.486	537,432			
Proxy Interview	Non-imputed sample							0.14	0.347	23,056
75th percentile of imputation	Non-imputed sample	0.202	0.401	15,728	0.154	0.361	537,432	0.198	0.399	23,056
90th percentile of imputation	Non-imputed sample	0.03	0.171	15,728	0.09	0.286	537,432	0.075	0.264	23,056
Potential attrition in the SIPP	Non-imputed sample							0.476	0.499	23,056
Received before reference period	True non-recipients	0.007	0.082	12,735	0.029	0.168	368,508	0.009	0.096	17,354
Received after reference period	True non-recipients	0.027	0.163	12,735						
Months since last receipt	True recipients	1.516	1.769	2,993	1.558	1.888	79,831	1.088	0.419	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
Months of receipt	True recipients	10.164	3.198	2,993	10.092	3.233	79,831	3.632	0.84	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

Notes: This table reports descriptive statistics for all the variables used in the Probit models. True non-recipients correspond to the sample used for the false positive analysis. They are classified as all households that we do not observe as participating according to the administrative data. True recipients correspond to the sample used for the false negative analysis. They are classified as all households that we observe as participating according to the administrative data. The non-imputed sample corresponds to the number of household to which question for participation in the Food Stamps Program was not imputed. Observations are weighted using survey weights adjusted for PIK probability using Inverse Probability Weighting.

Table 1.3: Descriptive Statistics for the Public Assistance Sample

		CPS			ACS			SIPP		
	Sample	Mean	SD	N	Mean	SD	N	Mean	SD	N
False positive rate	True non-recipients	0.006	0.078	17,156	0.016	0.126	526,566	0.005	0.073	24,066
False negative rate	True recipients	0.629	0.483	908	0.568	0.495	16,962	0.463	0.499	931
Survey response imputed	All sample	0.13	0.336	18,064	0.061	0.239	543,528	0.071	0.258	24,997
CATI Interview	Non-imputed sample	0.133	0.339	15,718	0.08	0.271	507,939			
CAPI interview	Non-imputed sample				0.389	0.487	507,939			
Proxy Interview	Non-imputed sample							0.14	0.347	23,069
75th percentile of imputation	Non-imputed sample	0.202	0.402	15,718	0.153	0.36	507,939	0.198	0.398	23,069
90th percentile of imputation	Non-imputed sample	0.03	0.17	15,718	0.065	0.246	507,939	0.075	0.264	23,069
Potential attrition in the SIPP	Non-imputed sample							0.476	0.499	23,069
Received before reference period	True non-recipients	0.005	0.07	14,939	0.024	0.154	401,086	0.004	0.061	20,634
Received after reference period	True non-recipients	0.009	0.092	14,939						
Months since last receipt	True recipients	2.533	2.926	779	2.308	2.939	14,610	1.196	0.615	840
Monthly amount received ($100)	True recipients	5.463	3.682	779	5.567	3.844	14,610	5.618	3.605	840
Months of receipt	True recipients	8.226	3.886	779	8.247	3.874	14,610	3.435	0.997	840
Participation rate in Zip Code	True recipients	0.141	0.094	659	0.134	0.094	14,570	0.09	0.062	831

Notes: This table reports descriptive statistics for all the variables used in the Probit models. True non-recipients correspond to the sample used for the false positive analysis. They are classified as all households that we do not observe as participating according to the administrative data. True recipients correspond to the sample used for the false negative analysis. They are classified as all households that we observe as participating according to the administrative data. The non-imputed sample corresponds to the number of household to which question for participation in the Public Assistance was not imputed. Observations are weighted using survey weights adjusted for PIK probability using Inverse Probability Weighting.

1.5 Results

To understand the determinants of measurement error, we construct different variables to test how survey design and respondent behavior affect errors in our data. We estimate a Probit model for each of the two error types as a function of different variables related to the theories of response error. We estimate specifications for each program and survey separately, as the variables available differ. For the false negative Probits, we restrict the sample to true recipients and estimate the determinants of the probability that they fail to report receipt. Similarly, the false positive estimates only include true non-recipients and we estimate the probability that they mistakenly report receipt. In each specification we include a set of demographic characteristics of the household: number of adults and children; sex, age, education, race, disability, and citizenship status of the household head; whether households are in rural areas, whether the household head speaks English poorly, report receipt of other programs; and a linear trend for calendar years in the survey. Observations are weighted using survey weights adjusted for the predicted linking probability using Inverse Probability Weighting. In Table 1.2 we present descriptive statistics for each variable we analyze for the FSP and in Table 1.3 for PA. We explain how we construct each variable as we describe the results, shown in Table 1.4 for the false negative responses and in Table 1.5 for the false positive responses.

1.5.1 The magnitude of errors in surveys

In the first rows of Table 1.2 and Table 1.3, we show the percentage of households classified as false negatives and false positives in each survey and program. There is wide variation in the error rates across surveys for the same program. For example, the false negative rate for the FSP in the CPS (42.4%) is more than two times the false negative rate in the SIPP (19.4%), while the false positive rate in the ACS for PA (1.6%) is three times or nearly so that of the SIPP (0.5%) and the CPS (0.6%). Since these three surveys are all random

samples of the same population, we might expect similar error rates, however, they are quite different. One likely reason for the differences is that survey design can substantially affect error rates. In fact, the three surveys we use differ along many dimensions. For instance, the main purpose of the CPS is to collect information on labor force status, while the SIPP is focused on gathering accurate information about participation in social programs. On the other hand, the ACS is a short-duration survey that respondents are legally obligated to answer, while the other two surveys are voluntary. In addition, the SIPP is a longitudinal survey in which respondents are visited every four months for four consecutive years, a higher interview burden compared to the other two surveys. Moreover, surveys impute program participation variables at different rates, which affects misclassification while it also reflects that respondent cooperation differs across the surveys.

Within each survey, the error rates are also different across programs. For example, the false negative rate for the PA program in the SIPP (46.3%) is two times that of the FSP program (19.4%), with similar differences observed in the other two surveys. The false positive rate also varies across programs within a survey; the FSP rate is more than triple that of PA in the CPS and the SIPP, while in the ACS false positive responses for PA are more frequent than for the FSP. The differences across programs in the error rates may reflect the fact that the pool of participants are different across programs or one program may be better known than the other.

The two error types in program participation reports may be explained by different factors. For instance, false negative responses may be the result of respondents' desires to speed-up the interview or their inability to recall events during the time frame of the survey. The false positive responses are less informative about misreporting errors in that its causes may be more related to data editing processes, such as imputation, or include households that recently moved from another state in which they received aid from the programs we

study.[32]

1.5.2 Determinants of misreporting program participation

We first test how imputation affects each error rate. We construct a binary indicator equal to one if the response to the question about program participation was imputed by the survey and equal to zero if the household responded to the question. The high imputation rates in some cases, shown in Table 1.2 and Table 1.3, already suggest that imputation could be a major source of error. For instance, the CPS imputes 13% of the sample in questions about FSP and PA participation. The ACS, has the lowest imputation rate in both programs, 1.1% in the FSP and 6.1% for PA, while the SIPP imputes approximately 7% of the sample in both cases.

The results in Table 1.4, show that imputation of a missing FSP response increases the likelihood of a false negative response in all surveys. In the CPS, the imputed responses are 24.2 percentage points more likely to be a false negative, while in the ACS the likelihood increases by 32.3 percentage points. In the SIPP, false negative responses increase 5 percentage points, though the effect is only significant at the 10% level. Again, the results are sizable with respect to the average false negative rate of the FSP in each sample. The probability of a false negative response is 1.5 times larger for imputed observations in the CPS and more than doubles in the ACS. Imputation of responses for PA participation increases the false negative rate in the CPS by 18 percentage points, for a relative effect of 29% with respect to the mean false negative rate in the sample. Across surveys and programs and types of error, the one case where the error is lower for imputed values is false negative reports of PA receipt in the ACS. In the ACS, imputed observations are 13.3 percentage points less likely to be a false negative, reducing the probability of a false negative rate by 23% relative to

32. Households that moved from another state recently may truly report participation 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.

the average. We discuss this further below, but the lower false negative rate comes at the expense of a very high false positive rate. The SIPP analysis shows no significant effects at conventional levels, possibly due to the large standard errors obtained after clustering.[33]

Table 1.4: Probit Estimates of the Determinants of False Negative Responses by Program and Survey (Marginal Effects)

	CPS	ACS	SIPP	CPS	ACS	SIPP
Survey response imputed	0.2421***	0.3229***	0.0498*	0.1798***	-0.1334***	0.0342
	(0.0193)	(0.0107)	(0.0266)	(0.0492)	(0.0116)	(0.0601)
Observations	3,539	81,772	4,771	908	16,962	931
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)
Monthly amount ($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)
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)
75th-90th imputation freq.	0.0556***	0.0100**	0.0063	-0.0236	0.0249**	0.0052
	(0.0201)	(0.0046)	(0.0272)	(0.0389)	(0.0121)	(0.0656)
90th-100th imputation freq.	0.3054***	0.0381***	-0.0084	-0.0198	0.0385**	-0.0116
	(0.0783)	(0.0052)	(0.0296)	(0.1479)	(0.0162)	(0.1001)
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)	
Proxy interview			0.0430			-0.0144
			(0.0271)			(0.0743)
Potential attrition in the SIPP			0.0084			0.0316
			(0.0225)			(0.0617)
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)
Non-Imputed observations	2,472	79,627	4,290	659	14,570	831
Average false negative rate:	0.421	0.257	0.194	0.629	0.568	0.463

Notes: This table reports marginal effects of different variables on the probability of a false negative rate in each survey and program. All regressions control for household composition, (composition of adults and children), sex, age, education, race, disability, and citizenship status of the household head, whether households are rural, speak English poorly, report receipt of other programs, and a linear trend for years of the survey. All models are estimated including all variables at once. 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<0.1

The results from the Probit models in Table 1.5, show that the false positive rate is

33. We estimate clustered standard errors using each household in the sample as a cluster to adjust for the correlation within households over time.

Table 1.5: Probit Estimates of the Determinants of False Positive Responses by Program and Survey (Marginal Effects)

	Food Stamps Program			Public Assistance		
	CPS	ACS	SIPP	CPS	ACS	SIPP
Survey response imputed	0.0276***	0.0313***	0.0175***	0.0112***	0.0232***	0.0068***
	(0.0027)	(0.0010)	(0.0029)	(0.0013)	(0.0006)	(0.0015)
Observations	14,525	461,756	20,226	17,156	526,566	24,066
Received before ref. period	0.0089	0.0072***	0.0199***	0.0012	0.0115***	0.0111***
	(0.0061)	(0.0009)	(0.0051)	(0.0018)	(0.0010)	(0.0026)
Received after ref. period	0.0061			0.0056***		
	(0.0045)			(0.0020)		
75th-90th imputation freq.	0.0045**	0.0011*	0.0047	0.0035***	-0.0005	0.0037**
	(0.0022)	(0.0006)	(0.0035)	(0.0010)	(0.0006)	(0.0016)
90th-100th imputation freq.	0.0046	0.0032***	0.0081	0.0022	0.0045***	-0.0001
	(0.0061)	(0.0007)	(0.0056)	(0.0027)	(0.0007)	(0.0015)
CATI interview		0.0017**			-0.0098***	
		(0.0007)			(0.0008)	
CAPI interview	0.0055*	0.0045***		0.0039**	-0.0095***	
	(0.0032)	(0.0006)		(0.0019)	(0.0005)	
Proxy interview			0.0038			-0.0010
			(0.0036)			(0.0014)
Attrition in the SIPP			0.0005			0.0017
			(0.0029)			(0.0012)
Non-Imputed observations	12,735	368,508	16,174	14,817	401,086	19,446
Average false positive rate:	0.020	0.012	0.015	0.006	0.016	0.005

Notes: This table reports marginal effects of different variables on the probability of a false positive rate in each survey and program. All regressions control for household composition, (composition of adults and children), sex, age, education, race, disability, and citizenship status of the household head, whether households are rural, speak English poorly, report receipt of other programs, and a linear trend for years of the survey. All models are estimated including all variables at once. 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<0.1

higher for imputed observations in all the surveys and programs we study. For example, false positive responses for participation in the FSP in the CPS are 2.8 percentage points higher for imputed observations. In the ACS the marginal effect of imputation on the false positive rate for the FSP is 3.1 percentage points, while for the SIPP, imputed observations are 1.8 percentage points more likely to be a false positive response in the FSP. For the PA program the effect of imputation is 1.1 percentage points in the CPS, 2.3 percentage points in the ACS, and 0.7 percentage points in the SIPP. Although the magnitudes may seem

small, they are sizable when compared to the average false positive rate in each sample, shown at the bottom of Table 5. The probability of a false positive response in the FSP almost doubles in the SIPP, more than doubles in the CPS, and more than triples in the ACS, with similar relative effects of imputation on the probability of a false positive PA response. The sizable relative effects of imputation reflect the fact that a large share of false positive responses is imputed in each survey.[34] Overall, the results show that imputation of missing data is a sizable component of both error rates, especially for false positive responses where imputation can more than double the probability.

There are other sources of both error rates as well. For instance, interviewers can make mistakes in the field, intentionally make-up responses, or respondents may be careless. We now focus on how other variables relate to each error rate within a survey. To do so, we exclude observations in which a given question about program participation is imputed. The main reason is that we want to study errors from misreporting rather than variation in errors induced by imputation during a post-processing stage of the survey.[35] All specifications are estimated including all variables at once and including the demographic controls.

The first variables we study are those related to the respondent's ability to place the events being surveyed in the specific time frames of each question. Even when households have the best intention to accurately answer a survey question, response errors may increase if respondents are unable to recall specific events or are confused about the timing of events. The three surveys we use record data of program participation using reference periods and we find that households tend to forget transfers that occurred a long time since the interview date and also find that people mistakenly associate receipt from other periods to the period in the survey, known as telescoping.

34. In the CPS the share of false positives that are imputed is 44.7% for the FSP and 50.4% for PA, while in the SIPP the share is 23.1% for the FSP and 27.6% for PA. In the ACS, 11.4% of the total false positive cases are imputed for FSP and 20.2% are imputed for PA.

35. Our main results do not change substantially, with the exception of survey mode. This is expected since item non-response rates vary substantially across modes of interview.

The first idea we test is whether respondents' ability to recall past events increases false negative responses. Households may forget and report no participation in government programs during the reference period if time went by since the last transfer and the interview is too long. In the administrative records, we are able to pin down the last calendar month that a household received a transfer within the reference period of each survey and test whether the number of months gone by since the last transfer and the time of the interview affect the rate of underreporting. The average number of months elapsed since the last transfer for each sample is shown in Table 1.2 and Table 1.3. Households that are true FSP recipients received their last transfer 1.5 and 1.6 months before the end of the reference period (12 months) in the CPS and the ACS. In the SIPP, true FSP recipients received the last transfer 1.1 months before the last month of the reference period (4 months). Households that are true PA recipients received their last transfer 2.5 and 2.3 months before the end of the reference period in the CPS and the ACS. In the SIPP, true PA households received the last transfer 1.2 months before the last month of the reference period.

The results in Table 1.4 show that for each month that goes by since the last FSP payment, households in the CPS are 3.6 percentage points more likely to underreport receipt, which corresponds to a relative effect of a 9% increase in the probability of underreporting FSP receipt. Likewise, each additional month elapsed increases the rate of a false negative error in the report of PA by 3.6 percentage points, which corresponds to a relative effect of 6% with respect to the sample average. The results for the ACS are similar in magnitude. Each additional month gone by since the last FSP receipt increases the probability of a false negative by 3.3 percentage points, which corresponds to 13% of the average false negative rate in the sample. For PA receipt, each month gone by since the last transfer increases the probability of a false negative by 3 percentage points, which corresponds to a relative effect of 5% with respect to the sample average. In the SIPP, the results show that the false negative rate of PA increases by 9.1 percentage points while the effect on reporting FSP

is not significant. For PA, the relative effect shows that the probability of a false negative response increases by 20% with respect to the sample average.

The second idea we test related to time frames in questions is that of telescoping effects and their effect on false positive responses. Reference periods may explain part of the false positive errors. For instance, transfers received before the reference period may be mistakenly brought forward in time because it takes effort to remember exact dates of participation. Likewise, recent events may be push backward into the reference period and increase the number of observations that report to have received aid in the survey but do not show up as recipients in the administrative records for the reference period.

In the CPS-ASEC, selected households are visited between February and April of a given year and are asked about participation in the past calendar year. In the ACS, each respondent is asked about her participation in the last twelve months, so the reference period is different depending on the month of the interview. In the SIPP the reference period is the last four months to the month of the interview. To test for whether telescoping effects explain false positive responses, we construct a binary indicator that equals to one if a household, that did not received a transfer within the reference period, participated in the program before the start of the reference period. For the CPS and the ACS we use twelve months before start of the reference period, and in the SIPP we use the last four months. Additionally, in the CPS, we construct a binary indicator that equals to one if the household received a transfer after the reference period and before the interview month (January - April of the year of the interview) and equal to zero otherwise. Table 1.2 and Table 1.3 show that 0.7% of true non-recipients received a transfer from the FSP in the previous twelve months of the start of the reference period in the CPS, while for PA the number is 0.4%. The numbers are 2.9% and 2.5% in the ACS, and 1% and 0.4% in the SIPP. The percentage of true non-recipients households in the CPS that received a transfer after the reference period is 2.7% for the FSP and 0.8% for PA.

The results in Table 1.5 show that receipt before the reference period has no significant effect on the likelihood of a false positive in the CPS, regardless of the program, however the point estimates are large with respect to the average false positive rate. In the ACS, we find that receipt of the FSP before the reference period increases the likelihood of a false positive response in 0.7 percentage points, which corresponds to a relative increase of 60% in the probability of a false positive response. For PA the telescoping effect shows that having received in months previous to the reference period increases the probability of a false positive response in 1.2 percentage points, which corresponds to two thirds of the average false positive rate. In the SIPP, receipt before the reference period increases the likelihood of mistakenly reporting receipt in 1.2 percentage points for the FSP and 1.1 percentage points for PA. Again, the effects are sizable with respect to the average false positive rates. Telescoping effects more than double and more than triple the probability of a false negative response in the FSP and PA in the SIPP. Finally, the results show that having received after the reference period in the CPS increases the likelihood of a false positive response for PA in 0.6 percentage points, which corresponds to almost doubling the probability with respect to the sample average. We find a similar but non-significant effect of backward telescoping for the FSP.

Another cognitive aspect that could affect response quality is that of the relative importance of government programs to households in the survey. Misreporting program participation may be lower if the topic, in this case government transfers, are more salient to a household. To measure saliency of program participation, we construct the total number of months a household received a transfer and the average monthly amount that they received during the reference period in each survey. The main idea is that households that are more dependent from government transfers should be less likely to make mistakes when reporting, holding other variables constant.[36] In the CPS and the ACS, the average number of months

36. Months of participation and monthly amount received are computed using administrative data so that they are actual months and amounts, not the ones reported. Since we observe these variables for those that

of participation is 10 months in the FSP and 8.2 months in PA within a twelve-month reference period (see Table 1.2 and Table 1.3). In the SIPP the average number of months of participation in a four-month reference period is 3.6 for the FSP and 3.4 for PA. The average monthly amount received is similar across surveys, approximately $300 USD in the FSP and $550 USD in PA (see Table 1.2 and Table 1.3).[37]

The results in Table 1.4 show that the false negative rate of the FSP in the CPS decreases by 1.5 percentage points for each additional month of receipt in the past calendar year. For PA, the results show that an additional month of participation reduces the false negative rate in 1.3 percentage points in the CPS. The results for the ACS show a similar effect of number of months. The probability of a false negative response for participation in the FSP decreases in 1.9 percentage points with an additional month of participation. The same analysis for PA shows that an additional month of participation reduces the false negative rate by 1.4 percentage points. We find similar effects of longer spells of participation in the FSP on the false negative rate in the SIPP. An additional month of participation in the FSP during the four-month reference period in the survey is associated to a 5 percentage points reduction in the false negative rate. We find no significant effects for months of participation in PA in the SIPP, plausibly because of the large standard errors we obtain after clustering.

We also study how the average amount received during the reference period affect the probability of misreporting. In most cases, higher amounts of receipt have a negative effect on the probability of underreporting, with the only exception of PA in the SIPP. Our results for misreporting FSP participation in the CPS show no significant effect, possibly due to a low estimated effect of amounts. For PA, we find that the probability of a false negative effect reduces in 1.1 percentage points for each $100 USD of monthly receipt. In other words, a household that receives a large transfer of $1000 USD, is 17% more likely to reveal true

are true recipients, we do not test a saliency effect on the false positive rate.

37. This is expected since all three surveys are intended to represent the same population.

participation in PA than a household that receives $100 USD. However, households that receive such a large amount are rare.[38] As such, the effect of monthly transfers is significant but relatively small. This is similar across other surveys and programs. In the ACS, the likelihood of reporting FSP decreases by 0.3 percentage points for each $100 USD, while for the results for PA show no significant effect, possibly due to a low estimated effect of amounts. In the SIPP, the false negative rate in the FSP decreases by 1.2 percentage points per $100 USD received. The only case where we do not see a positive relation between underreporting and amount received is that of PA in the SIPP, where the probability of a false negative response increases by 1.5 percentage point with each $100 USD received. Overall, our results show that more salient events are reported better, however, the effects are small when compared to other determinants of the false negative responses.

The next variables we study are related to how households' cooperation or "tastes" for surveys relate to misreporting. We construct a measure for survey cooperation as the frequency in which different questions of the survey are imputed. Households with a higher imputation frequency in other sections of the survey are expected to be households that are less cooperative or households that dislike survey participation the most and misreport at a higher rate. We select variables that are asked to every household in each survey and compute the percentage of questions that are imputed within the set of variables selected. Since the number imputed answers varies across surveys, and since each survey may use different criteria for the decision of imputing missing data, we standardized our measure of cooperation. In particular, we divide the distribution of the imputation frequency in three categories: i) households above the 90th percentile, ii) households above the 75th percentile and below the 90th percentile, and iii) households below the 75th percentile in the distribution of imputed answers within a survey. We include the highest and second highest categories in the probability models and leave households below the 75th percentile

38. For instance, in the case of FSP (SNAP) a household with 7 or more members and no income deductions receives above $1000 USD. See $http://otda.ny.gov/programs/snap/$.

as the base group. With three categories we can test for a non-monotonic relation between households cooperation with surveys and the probability of misreporting answers. Table 1.2 shows that 20.2% of the CPS sample is in the 75th-90th percentile category, while 3% of the sample is above the 90th percentile in the distribution of imputed variables. The remainder of the sample is below the 75th percentile. In the ACS, 15.4% of the sample is in the 75th-90th percentile category, while 9% is above the 90th percentile of the imputation distribution. In the SIPP, 19.8% of the sample is in the 75th-90th percentile category, while 7.5% of the sample is above the 90th percentile of the imputation distribution in the survey.

The results in Table 1.4 show a consistent non-monotonic relation between the false negative rate and survey cooperation. For instance, CPS households in the 75th-90th percentile category are 5.6 percentage points more likely to not report receipt of FSP, while households that are imputed most frequently are 31 percentage points more likely to underreport receipt. Taken together, the average probability of misreporting true receipt of the FSP almost doubles for households in the upper quartile of the imputation distribution. Furthermore, households above the 90th percentile are six times more likely to misreport true receipt of the FSP than households above the 75th percentile. The analysis of false negative responses for PA in the CPS shows small and non-significant effects. A similar non-monotonic relation between survey cooperation and underreporting is shown in the ACS for both programs. The false negative probability in the report of participation in the FSP increases in a percentage point for households in the 75th-90th percentile category, while for households above the 90th percentile the probability of a false negative response is 3.8 percentage points higher with respect to households below the 75th percentile. Taken together the probability of not reporting participation in the FSP is 17% higher for households in the upper quartile of the imputation distribution. The results are similar when we study report of PA participation. Households in the 75th-90th percentile category are 2.5 percentage points more likely to not reveal true participation in PA compared to households below the 75th percentile. The

marginal effect is 3.9 percentage points for households above the 90th percentiles. Combined, households in the upper quartile of the imputation frequency in the ACS are 11% more likely to not report participation in PA.

The results in Table 1.5 show that the false positive response in the report of participation in the FSP increases by 0.1 percentage points and 0.32 percentage points for observations in the 75th-90th percentile category and above the 90th percentile in the ACS, compared to more cooperative households classified below the 75th percentile of the imputation distribution. Adding both marginal effects, the probability of a false positive response increases by 35% for households in the upper quartile of the imputation distribution. For PA, we find a significant increase in 0.5 percentage points in the false positive rate for the group of less cooperative households, while the marginal effect of households in the 75th-90th percentile category is small and not significant. Relative to the average, the probability of a false positive response in the report of PA is 28% higher for households above the 90th percentile of the imputation frequency in the ACS. Our results for the CPS show that the false positive response of participation in the FSP increases in 0.5 percentage points for households in the 75th-90th percentile category of the imputation frequency. In other words, households in this category are 23% more likely to give a false positive response than households below the 75th percentile. When we study the report of PA the results show that the false positive response increases in 0.4 percentage points for households in the 75th-90th percentile category, which corresponds to a 58% increase in the probability of false positive response relative to the average. The analysis in the SIPP shows a similar relation between survey imputation and error rates, however in most cases the marginal effects are estimated imprecisely. Households in the 75th-90th percentile category are 0.4 percentage points more likely to give a false positive response compared to households below the 75th percentile, which corresponds to a 74% increase with respect to the average probability of a false positive response in the survey. Overall, our results show that proxies for non-response profiles within a survey are

related to misreporting, especially in the case of false negative responses.

We next explore how response errors relate to the interview mode used during the survey. Interview mode effects will have a different interpretation depending on the survey since different modes are used for different reasons. For instance, the ACS sends a questionnaire by mail to every household selected to participate in the survey (see US Census Bureau, 2014). After several attempts, households that fail to send back a completed survey are contacted by telephone. If telephone interviews are not successful or cannot be conducted, households are visited for an in-person interview. As such, the mode of interview applied is likely related to how reluctant to participate in the survey is a selected ACS household. In the CPS, on the other hand, uses telephone or in-person interviews. Cases are assigned to telephone interviews if i) households have a telephone and are willing to accept a telephone interview, ii) the field representative recommends a telephone interview, or iii) if the interview month is not the first nor the fifth interview of the household (US Census Bureau, 2006). As such, whether a case is interviewed by telephone or in-person is determined in part randomly, according to the interview month of the household, and in part by the decision of a household to be contacted by telephone.

To test for the relation between interview modes and response errors in the ACS, we construct a binary indicator for whether the interview was administered by telephone or in-person, and compare how false positive and false negative responses differ from those in mail interviews. Table 1.2 and Table 1.3 show that 38.2% of the non-imputed sample is interviewed in-person (CAPI) while 7.8% of the sample is interviewed by telephone interviews (CATI). The remainder of the sample is self-administered by mail. In the CPS, we construct a binary indicator for whether the interview was administered by telephone. Table 1.2 and Table 1.3 show that 12.7% of the sample is interviewed by telephone (CATI) while the rest is interviewed in-person. We are not able to explore interview mode effects in the SIPP since it only conducts in-person interviews.

The results for the ACS in Table 1.4 shows that telephone interviews are on average 8.1 percentage points more likely to be a false negative report than mail interviews, while face-to-face interviews are 16.4 percentage points more likely to not report receipt of the FSP than self-administered surveys. In terms of the average false negative responses in the ACS, the results show that the probability almost doubles for the FSP in non-mail surveys, after adding both coefficients. The marginal effects of survey mode are similar for PA in the ACS. Telephone interviews are 6.6 percentage points more likely to give a false positive response compared to mail surveys, while in-person interviews are 14.3 percentage point more likely to do so. The relative effect of non-mail interviews, after adding both coefficients, shows that the probability of a false negative response increases by 37% in non-mail surveys relative to the average error. Furthermore, in both programs the results in the ACS show that the marginal effect of in-person interviews is twice that of telephone interviews. Likewise, in the CPS the false negative rate in the report of the FSP increases in 15.3 percentage points in face-to-face interviews as opposed to telephone interviews. Relative to the average false negative rate in this sample, the relative increase is 36% in the probability of a false negative response in interviews conducted in-person as opposed to by telephone. The marginal effect of in-person interview on the false negative response for PA is not significant and small relative to the other cases we find. While in both surveys we find that telephone interviews have lower error rates than in-person interviews, the differences in error rates between telephone and in-person interviews in the ACS can also be interpreted as differences between more and less cooperative households. Less cooperative households are more likely to be interviewed in-person, and they are also more likely to give worse responses.

Our findings in Table 1.5 show that telephone interviews are on average 0.2 percentage points more likely to be a false positive in the FSP compared to self-administered mail interviews in the ACS. Likewise, in-person interviews are 0.5 percentage points more likely than mail interviews to mistakenly report receipt of the FSP in the same survey. Adding

both coefficients, the probability of a false positive response in the FSP increases by more than 50% for non-mail interviews when compared to the sample average. For cash welfare we find the opposite association between survey mode and false positive responses. The false positive rate of reporting PA is a percentage point lower in telephone surveys and face-to-face interviews when compared to surveys responded by mail. In other words, the probability of a false positive response in PA is three times higher in self-administered surveys. The results for the CPS show that face-to-face interviews are, on average, 0.5 percentage points and 0.4 percentage points more likely to mistakenly report receipt of the FSP and PA, when compared to telephone interviews. Relative to the average false positive rate in the sample, the probability increases by 25% for the FSP for in-person interviews, and 67% for PA.

Another possible variable related to households' cooperation is whether information for some members of the households was gathered through proxy interviewing. Using the SIPP sample we construct a binary indicator that equals to one if the interview for the reference person of the household was done using proxy interviews, and equal to zero otherwise. Approximately 14% of the observations we use in the SIPP obtain information from the reference person through proxy interviews. The results for the false negative rate in Table 1.4 and false positive rate in Table 1.5 show no significant effects of proxy interviewing on the error probabilities. Finally, we take advantage of the longitudinal structure of the SIPP and construct a binary indicator that equals to one if a household "attrits", i.e. leaves the sample in any of the subsequent waves; or equal to one if a household rejects to answer the survey in any wave. The fraction of the sample that "attrits" or rejects to answer any wave in the next period of the survey is 48%. The results in Table 1.4 and Table 1.5 show a positive but small and non-significant correlation between future attrition and the probability of misreporting.

Finally we study how misreporting true participation is related to stigma associated to participation in government programs. The main idea we want to test is whether households for which stigma of participating in cash welfare or the FSP is higher are less likely to reveal

participation in a program to an interviewer. We assume that welfare stigma should be lower in areas where there are higher rates of participation. In other words, social desirability effects of program participation in areas of higher participation should point towards welfare participation being a more acceptable, hence less embarrassing, behavior and be more likely to be revealed to a third person. For each household, we construct the proportion of households that participate in each program in their ZIP-Code of residence. The participation rate is constructed as the ratio of total cases within a ZIP-Code obtained from the administrative data, to total housing units in the same area obtained from the US Census 2010. We compute the yearly participation rate and assign it to households interviewed in the same year. The average participation rate in the FSP at the ZIP-Code level is 51% in the CPS, 48% in the ACS, and 34% in the SIPP (see Table 1.2). Likewise, for the PA program the average participation rate at the ZIP-Code level is 14% in the CPS, 13% in the ACS, and 9% in the SIPP (see Table 1.3).

Our results in Table 1.4 show that an increase of 10 percentage points in the FSP participation rate in the ZIP-Code of residence is associated with a decrease of 0.8 percentage points in the probability of misreporting true program participation in the CPS. The results show an elasticity of ZIP-Code participation and false negative responses of 0.09, i.e. a one-percent increase in ZIP-Code participation reduces the false negative rate in 0.09%. For PA, the results show that a 10 percentage point increase in the PA participation rate in the ZIP-Code of residence is associated with a decrease of 3.2 percentage points in the probability of misreporting true participation. The implied elasticity of ZIP-Code participation and the probability of a false negative response in PA is 0.07, i.e. a one-percent increase in ZIP-Code participation reduces the false negative rate in 0.07%. In the ACS a 10 percentage-point increase in PA participation in the ZIP-Code reduces the likelihood of underreporting PA in 1.3 percentage points, for an elasticity of 0.03. We find no significant effect of ZIP-Code participation in the FSP and the probability of false negative responses

in the ACS, plausibly due to the small magnitude. In the SIPP, the point estimates are large and positive, but very imprecise, plausibly due to clustering standard errors.[39]

1.5.3 Robustness to the presence of interviewers

The effect of respondents' characteristics on response errors may also depend on how well interviewers perform in the field. For instance, interviewers may invest higher efforts in poorer areas to reduce false negatives in the report of cash welfare or FSP participation. Interviewers may also help in any clarification that respondents need while answering a survey. Moreover, households that refuse to answer the survey by telephone and are subsequently visited by an interviewer may differ substantially from households that agree to be interviewed by telephone or answer the survey by mail. Ideally, one would have data on interviewer characteristics and test how our results change with one interviewer to another. While we do not have such variables, we are able to test whether our results are different for households that respond the survey with an interviewer to households that respond without any interaction with interviewers.

In particular, we use the linked data from the ACS to test whether our results are robust across mail interviews and non-mail interviews. Given the large sample size in the ACS, the mail and non-mail subsamples are still large enough to do a separate analysis within them. We estimate the exact same models in each subsample.

The results in Table 1.6 show the results for months elapsed since last receipt on the rate of false negatives remains similar across interview modes in the ACS. When we look at the telescoping effects on the false positive rate of the FSP, the effects are very similar between mail and non-mail interviews. The results for PA show that the telescoping effects almost double in self-administered surveys when compared to non-mail interviews. The differences across survey modes may reflect that interviewers are useful in clarifying questions about

39. In the SIPP sample, we end up with a small number of households so that variation accross ZIP-Code in the rate of participation is even lower after aggregating our data.

which programs qualify under the definition of welfare programs in the ACS. The effects of saliency are also very similar across modes of interview.

The relation between our measures of cooperation and the error rates is mixed and are mostly case specific. For instance, point estimates of the effect of survey imputation on the false positive rate are very similar in mail and non-mail cases, but only significant for non-mail, which probably reflects that the sample of non-mail surveys almost triples that of mail surveys. With the only exception of false positive responses of PA for the non-mail sample, our results show that less cooperative households are more likely to misreport, with no clear differences across type of interview.

Finally, ZIP-code level participation in the FSP has a negative and significant effect when respondents are interacting with interviewers in the ACS. A 10 percentage-point increase in the FSP participation in the ZIP-Code reduces the likelihood of underreporting FSP in 0.22 percentage points. The effect of ZIP-Code participation on the probability of a false negative response of participation in PA participation is only significant for non-mail interviews.

Table 1.6: Probit Estimates of the Determinants of Errors in the ACS by Program and Interview Mode (Marginal Effects)

	False Positive Response				False Negative Response			
	Food Stamps Program		Public Assistance		Food Stamps 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)
Received before ref. period	0.0076***	0.0073***	0.0080***	0.0156***				
	(0.0018)	(0.0007)	(0.0013)	(0.0013)				
Monthly amount ($100)					-0.0029*	-0.0031**	-0.0027*	0.0008
					(0.0016)	(0.0012)	(0.0014)	(0.0019)
Months of receipt					-0.0193***	-0.0172***	-0.0147***	-0.0115***
					(0.0008)	(0.0005)	(0.0016)	(0.0020)
75th-90th imputation freq.	0.0028	0.0024***	-0.0023**	0.0016**	0.0115*	0.0001	0.0133	0.0252
	(0.0020)	(0.0005)	(0.0011)	(0.0007)	(0.0069)	(0.0045)	(0.0159)	(0.0168)
90th-100th imputation freq.	0.0016	0.0010**	0.0084***	0.0047***	0.0654***	0.0110**	0.0089	0.0494***
	(0.0014)	(0.0005)	(0.0018)	(0.0008)	(0.0112)	(0.0043)	(0.0342)	(0.0183)
Participation rate in Zip Code					-0.0221**	0.0214***	-0.1419**	-0.0310
					(0.0096)	(0.0061)	(0.0636)	(0.0810)
Non-Imputed observations	99,147	269,361	119,985	281,101	37,270	42,357	8,299	6,271

Notes: This table reports marginal effects of different variables on the probability of a false positive and a false negative rate in the ACS for each and program. We divide the analysis into the sample of households that were interviewed by mail and households that were interviewed by non-mail interviews (CATI or CAPI). All regressions control for household composition, (composition of adults and children), sex, age, education, race, disability, and citizenship status of the household head, whether households are rural, speak English poorly, report receipt of other programs, and a linear trend for years of the survey. 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<0.1

The results are consistent with the fact that stigma should only matter when program participants are revealing participation to another person, to which respondents feel or at least have some belief of being judged.

1.6 Discussion

Our results have a number of suggestions for users of survey data. Imputation of missing responses is a large source of survey errors, and in most cases increases error rates substantially. The only exception is the case of PA receipt in the ACS where imputation reduces the false negative error, though at the expense of inducing a large false positive error. A common discussion to decide whether imputed values in survey data should be used or not is based around the assumption that observations are missing at random after we account for observable differences between respondents and non-respondents of a particular question (see Rubin, 1976). This is usually known as the missing-conditionally-at-random assumption (MCAR). The main arguments for or against MCAR have to do with how well does the method implemented to impute missing data solves for selective item non-response. Our results add evidence to this discussion. In particular, even if researchers are willing to accept the MCAR assumption, embedded in most imputation methods, our results show that imputation of missing data by surveys leads to substantially higher error rates. To the extent that the errors generate bias, which has been shown by others, imputation may be a driver of bias from measurement error in survey data. In fact, Figure 1 shows that imputation of missing data is responsible for 18% of the bias in total dollars reported from the FSP when using the CPS data. Researchers could do better if they work with the subsample of non-imputed households and use their own methods to control for selective non-response.[40] This is in line with what others have recommended when studying the effect of imputed

40. One alternative would be to control for non-response using Inverse Probability Weighting (see Wooldridge, 2007) so that researchers work with a sample that is free of error due to imputation while still attempt to control for selective non-response.

values of earnings on descriptive analysis of the income distribution or when estimating the returns of schooling on earnings (see Lillard et al., 1986 and Hirsch and Schumacher, 2004).

Another implication is that scientists can no longer presume that nonresponse and response errors are separable. Both confound model estimation. Conventional weighting schemes for nonresponse will often not improve model estimates. In panel data it may be that nonresponse profiles can proxy for some part of response error. The extent to which that proxy assists estimation can only be determined from careful validation of a substantial number of domains of survey measurement. Should our hypothesis be supported in many studies of validation data, survey design would need to redirect resources that are now directed at reducing nonresponse toward measuring response errors.

Our results also show that errors in response are lower for households with longer spells and that receive higher amounts. Households who are more dependent on government transfers report better, which could be appealing for researchers that use survey data to study the effects of government programs or study measures of well-being for people at the bottom of the income distribution. However, the magnitudes are small. The most extreme case compares households that participate for a single month to households that participate for the full length of the reference period. The relative increase in the probability of a false negative response, between a single-spell household and full-spell household, ranges from 23% in PA in the SIPP to 81% in the FSP in the ACS. The false negative rate decreases substantially; however, cases that participate for one-month are rare. Using values of the distribution of months of participation in the CPS, the probability of a false negative response increases by 21% for household above the 80th percentile of participation months (12 months) compared to those below the 20th percentile (6 months). This corresponds to about half of the effect comparing full spell to single-spell households. In addition, other components of the false negative rate have much larger effects than our measures of salience. For instance, consider the effects of recall where we look at how months gone by between the interview date and

the last transfer affect the false negative rate. Comparing a household for which 12 months have gone by with a household that was interviewed just one-month after the last transfer received, the results show that the probability of a false negative response doubles in the CPS. The relative effect of recall corresponds to two times the relative effect of participation months on the rate of false negative responses. In other words, the gains in response quality that we obtain from salience effects, even in such extreme cases, compensate for half of the loss in response quality obtained from recall effects, which is only one of the components that increase the rate of false negative response. The results are similar for other surveys and programs.

Our results also have a number of suggestions to survey producers. Survey agencies, as well as researchers that gather their own data, are in most cases worried about overall non-response rates of the surveys they make. This is reflected by the fact that survey non-response has been the main topic of attention for decades and most efforts are dedicated to reducing non-response rates (Massey and Tourangeau, 2013). However, households that are more reluctant to participate in surveys or have a high opportunity cost of time may also be households that, once they agree to participate, are more likely to misreport than others (see Krosnick, 1991 and Tourangeau et al., 2010). To the extent that such behavior is proxy by our measures of cooperation, our results support the hypothesis of a trade-off between increasing response rates and improving survey quality.

Survey errors are also the result of how questions are designed in surveys. The analyses of the effects of reference periods show that time frames play an important role in explaining errors in the report of program participation. We find that longer periods of recall increase the probability of misreporting true participation. We also find strong evidence that telescoping effects are an important component of the false positive rate. As such, researches need to be aware that they may be inducing recall or telescoping effects in the response when they design the survey questionnaire. Some suggestions to reduce recall and telescoping effects

are available in the survey literature. For instance, Sudman and Bradburn (1973) suggest using references to important events that occurred around reference periods instead of referring to particular dates. However, validation data for the sample or a random sub-sample of participants in the survey could be used to better assess how strong are the effects of time frames in questions.

Our results also show that the presence of interviewers can influence responses. Interviewers may improve responses. For instance, the presence of interviewers in the ACS increases the false positive rate in the FSP while it reduces the false positive rate of PA. The differences across the two programs in the ACS may reflect the fact that the question for cash welfare in the ACS includes a broad list of programs, from federal to state or local welfare offices, which may be confusing to respondents. Interviewers may help to clarify any confusion in telephone or in-person interviews and reduce the over-reporting in both of these modes relative to mail interviews. However, the results in the ACS also show that there is higher under-reporting when an interviewer is involved. Furthermore, the CPS and ACS show higher error rates for in-person interviews compared to telephone interviews, which suggests that the physical presence of an interviewer increases the likelihood of misreporting.

Likewise, our results of stigma show that respondents tend to underreport socially undesirable behavior, and that the stigma effect is more evident when an interviewer is present. Answering surveys may threaten the respondent in different ways. They may fear losing part of their benefits if their information about participation in social programs is made public to local Department of Social Services, or may fear being morally judged by the interviewer or others in the community. [41][42]

41. See Heffetz and Ligett (2014) for a discussion on privacy and anonymization in data-based research.

42. In fact people often judge those that live off government transfers since there is a social norm installed that pressures people towards economic self-sufficiency. This has been studied as the tax payer resentment by Besley and Coate (1992) and more recently by Lindbeck et al. (1999).

1.7 Conclusion

In this paper we analyze different causes of measurement error in the report of government transfers. We link the ACS, the CPS, and the SIPP to administrative micro-data of FSP, TANF, and General Assistance transfers from New York State. We study two types of errors in binary responses of program participation: false negative responses and false positive responses.

While there are many studies that explore causes of measurement error 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 missing much of the details found in micro-data; and studies using micro-data generally use one survey, study a single program, or use data that is 30 years old. We provide a more consistent picture by comparing our results across different surveys and within surveys between different variables. Our results have relevant implications to a broad area of research in economics that use or make their own surveys.

Error rates are different across surveys and programs, which likely relates to how surveys differ in their design and data post-processing methods. We find that false positive responses are largely explained by imputation of missing data and false negative responses increase with imputation in most cases. Our results also suggest that error rates can be the result of having a group of households that are non-cooperative and misreport as part of a pattern of low effort. Efforts invested in convincing households to participate in the survey may be screen less cooperative households into the survey, which suggest a trade-off between increasing response rates in the survey and improving the quality of the data. Our findings also confirm other behavioral theories of why people misreport in surveys. We find that response errors are related to respondents' ability to recall specific events and to place them in a particular timeframe (telescoping). Our results also show that salience of the topic improves the quality of the answer. We also find that households tend to under-report

socially undesirable behavior.

Our results and recommendations are broad enough to be applied in many examples where researchers believe that measurement error 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 creative methods to control for misreporting in their own data. In any case, the best alternative is to invest efforts in linking survey data to administrative records, or other reliable sources of information.

CHAPTER 2

LONG-TERM EFFECTS OF TEMPORARY INCENTIVES ON MEDICAL CARE PRODUCTIVITY

*In collaboration with Paul J. Gertler, Paula Giovagnoli, and Christel Vermeersch**

2.1 Introduction

A well-documented feature of technological change is its remarkably slow diffusion.[1] One reason could be that innovation can be costly above and beyond acquisition expenditures. Firms have to design, test, and learn how to best incorporate an innovation into existing practices. Firms may also need to purchase complementary technology (Rosenberg, 1982; David, 1990; Bresnahan and Trajtenberg, 1995). Productivity might also be lower during a period of adjustment while the firm implements and learns how best to use the innovation. Management may have to overcome costly informational deficits (Bloom et al., 2012) and behavioral barriers such as present bias (Duflo et al., 2011). Moreover, innovation may confront worker's resistance if part of their wage varies with performance (Lazonick, 1979; Atkin et al., 2015) or if they have developed strong work habits. These fixed costs of adjustment can be large enough to prevent productive and profitable innovation.[2] Change is hard, and even small fixed costs may inhibit changes in favor of maintaining the status quo (DellaVigna, 2009; Thaler and Sunstein, 2009).

∗. Gertler: Li Ka Shing Professor of Economics at the University of California, Berkeley (email: gertler [at] haas.berkeley.edu); Giovagnoli: Economist at The World Bank (email: pgiovagnoli [at] worldbank.org); Vermeersch: Senior Economist at The World Bank (cvermeersch [at] worldbank.org).

1. Slow adoption of new technologies by firms has been extensively documented in agriculture, manufacturing and medicine. Surveys and studies of slow diffusion include Acland and Levy (2015), Ryan and Gross (1943), Griliches (1957), Mansfield (1961), Coleman et al. (1966), Rosenberg (1972), Parente and Prescott (1994), Foster and Rosenzweig (1995), Geroski (2000), Hall and Khan (2003), and Comin and Hobiijn (2010), Conley and Udry (2010).

2. The organizational literature refers to the phenomenon of high fixed costs of preventing the adopting profitable innovations as organizational inertia (Hannan and Freeman, 1984; Carroll and Hannan, 2000).

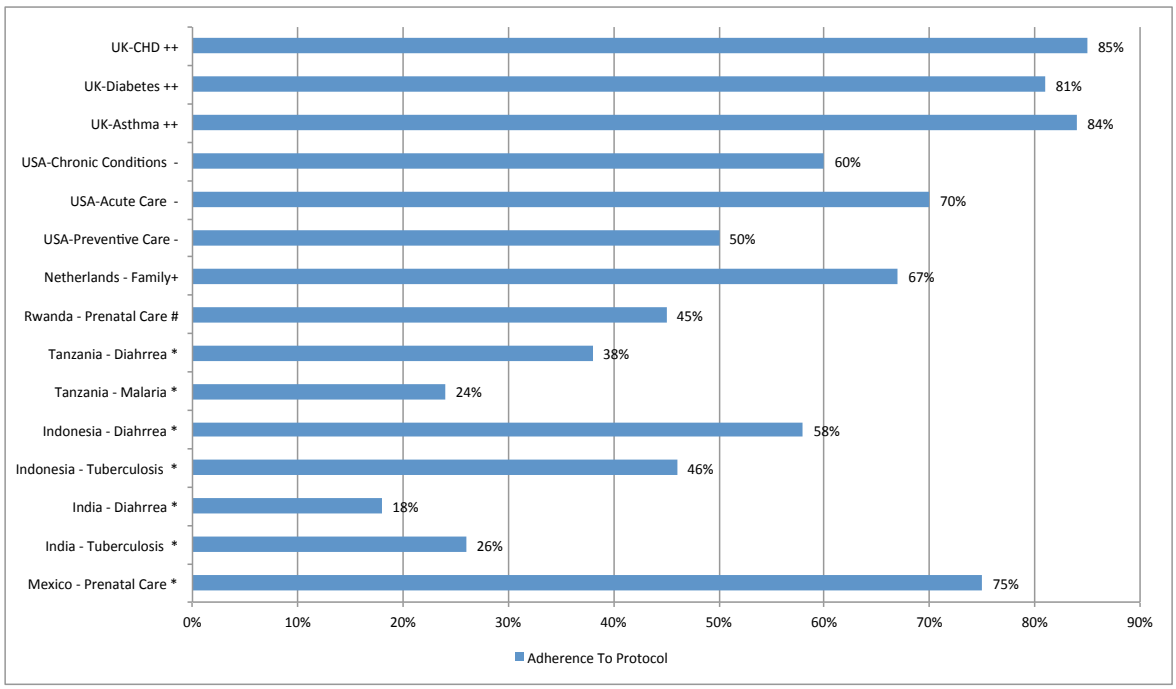

Figure 2.1: Provider Compliance with Clinical Practice Gudelines

Notes: Authors' elaboration based on (-) Schuster et al. (1998); (+) Grol (2001); (++) Campbell et al. (2007); (*) Das and Gertler (2007); and (#) Gertler and Vermeersch (2012)

Slow diffusion of better quality medical care practices is a global issue as evidenced by the remarkably low level of compliance with Clinical Practice Guidelines (CPGs) (Figure 2.1).[3] While low CPG compliance may in part reflect a lack of knowledge or slow diffusion of information (Phelps, 2000), evidence shows that practitioners often provide a standard of care well below their level of knowledge of CPGs.[4] In a systematic review of the literature, Cabana et al. (1999) report that resistance to changing existing practice patterns is one of

3. CPGs define medical care production possibility frontiers in that they prescribe the clinical content of care that maximizes the likelihood of successful health outcomes based on medical science, clinical trials, and practitioner consensus. Local CPGs are regularly updated and serve ainformations the basis of training in medical schools and practitioner refresher courses.

4. See Das and Hammer (2005), Das and Gertler (2007), Das et al. (2008), Barber and Gertler (2009), Leonard and Masatu (2010), Gertler and Vermeersch (2012), and Mohanan et al. (2015).

the most important barriers to CPG adherence.[5] For example, Grol and Grimshaw (2003) report that 49% of UK nurses and doctors said that resistance to changing old habits was a major obstacle to complying with new hand hygiene guidelines. In a recent study of hospitals in the U.S., Skinner and Staiger (2015) state that the large differences observed in the adoption of aspirins by providers to treat heart disease conditions may be in part due to resistance to changing old habits of physicians.[6]

In this paper we examine the short-term and long-term effects of paying temporary financial incentives to medical care providers in Argentina to increase the share of women who initiated prenatal care in their first trimester of pregnancy. Before the intervention the share of women who initiated care early was low and medical care clinics had no specific activities or practices devoted to identifying and encouraging pregnant women to start care early. Providers began prenatal care whenever women choose to first show up at the clinic. We show that the temporary incentives motivated clinics to invest time and effort to develop and test new data driven methods of how to best identify newly pregnant women and encourage them to seek care early. In other words, clinics innovated in the sense that they had to experiment with new outreach strategies until they learned what worked best.

How best to encourage better quality care depends on whether slow adoption is due to low perceived value or to high fixed costs of adoption. If providers value a new clinical service but are reluctant to adopt because of high fixed costs, then a temporary incentive large enough to cover these costs should lead providers to adopt the service and continue it after the incentive is removed.[7] If on the other hand, fixed costs are low but providers have a

5. For more evidence of resistance to change as a barrier to CPG compliance see Grol (1990), Hudak et al. (1995), Main et al. (1995), and Pathman et al. (1996).

6. For other references on technology adoption in the medical care industry see Baker (2001), and Baker and Phibbs (2002), Berwick (2003) Cutler and Huckman (2003), Cutler (2007), and Bech et al. (2009).

7. In practice, this amounts to paying providers a time-limited per unit incentive for the new service. Paying an upfront lump sum amount is another option. However, it may be harder to ensure and verify the actual change in practice patterns. By paying based on actual performance the incentives also include a commitment device for compliance.

low perceived value of the service, then the price increase should also lead to adoption while the increase is active, but the service will be dropped once the incentive disappears. Hence, studying provider behavior both while incentives are active and after the incentives are removed provides a test of whether fixed costs versus low returns are inhibiting innovation.

Early initiation of prenatal care has long been part of the Argentine CPGs for prenatal care, and is part of standard training in Argentine medical and nursing schools as well as throughout the world (Organization, 2006). Providers are taught that prenatal care by skilled health professionals beginning in the first trimester of pregnancy is essential for good maternal and newborn health outcomes as it is argued in the medical literature (Schwarcz et al., 2001; Carroli et al., 2001; Campbell and Graham, 2006a). Through early initiation of care, providers are able to detect and correct important medical conditions such as maternal infections or anemia in the period in which the fetus is most at risk and before these conditions can jeopardize maternal or newborn outcomes (Carroli et al., 2001; Hawkes et al., 2013). Early prenatal care also allows providers to advise mothers on proper prenatal nutrition and prevention activities in the period in which the fetus is developing most rapidly.

We use data from a field experiment conducted with *Plan Nacer*, an Argentine government program similar to Medicaid in the U.S. and *Seguro Popular* in Mexico that provides health insurance to otherwise uninsured pregnant women and children.[8] The field experiment randomized temporary financial incentives to health care clinics in which treatment clinics were paid a 200% premium for the first prenatal care visit if the visit occurred before the 13th week of pregnancy. The fee increase was paid for 8 months and then removed. Clinics were explicitly made aware that the fee was temporary.

We find that the rate of early initiation of prenatal care was 34% higher in the treatment group than in the control group (0.42 versus 0.31) and that the average weeks pregnant at the time of the first prenatal care visit fell by about 1.5 weeks while the incentives were

8. In 2013, *Plan Nacer* was expanded to other populations and renamed *Programa Sumar*.

being paid. We then show that that the higher levels of early initiation of prenatal care in the treatment group persisted for at least 24 months after the incentives ended.

We also document that clinics developed specific data driven strategies to find newly pregnant women and encourage them to start care early. Clinics designed and tested new beneficiary outreach strategies such as (i) coordinating with local pharmacies to keep track of women who stopped using birth control pills, (ii) meeting with teenagers while parents were less likely to be at home so that they would be more prone to reveal pregnancies, (iii) talking with mothers when they come to pick free milk for children and (iv) modifying gynecologist schedules to be able to more easily make appointments. These new strategies took time to develop and test, and involved opportunity costs to clinical staff beyond marginal costs of actual implementation. We show that all outreach activities doubled in the treatment group relative to the control group during the intervention period, and that this increase persisted at least 15 months after the incentives ended.

We also provide evidence that it was in the clinics' interest to have provided these outreach services absent the fixed costs of adjustment. First, in a survey discussed later, clinic medical directors reported that early initiation of care was ranked as one of the highest of health priorities among all prenatal care services. Second, *Plan Nacer* reimbursed clinics for beneficiary outreach activities at a rate higher than the cost of delivering those activities. Finally, prenatal care visits by *Plan Nacer* beneficiaries were profitable to clinic staff since 50% of the fees obtained from prenatal visits were used to pay wage bonuses.

Taken together these results are consistent with the presence of fixed costs as opposed to low perceived value inhibiting the diffusion of better quality of care practices. First, early initiation of care was both profitable at the lower fees and perceived to be important for health outcomes; yet, absent the incentives it was being provided at a low rate. Second, temporary incentives lead to the development of new outreach activities designed to identify and encourage pregnant women to seek care early resulting in a large and significant increase

in the early initiation of care that persisted long after after the incentives ended. Third, the new outreach activities were in and of themselves profitable.

Our results suggest that fixed costs of adoption maybe the mechanism behind recent evidence that permanent performance incentives improve access to higher quality medical care.[9] The standard explanation is that providers are reallocating their effort across services in response to the increased profit opportunities.[10] However, previous studies have been unable to distinguish between this mechanism and a fixed cost of adoption story. We are able to distinguish between the two mechanisms by observing what happens when incentives are removed. While the incentives are in play both models predict a positive response. However, once the incentives are removed, practice patterns should revert to prior levels in the standard models but continue at the higher levels under the fixed costs of adoption model. Understanding the mechanism by which financial incentives work is policy relevant. If temporary financial incentives are able to induce providers to adopt permanent changes to their clinical practice patterns, then temporary incentives can achieve a long-term boost in performance cheaper cost than permanent incentives.

Temporary incentives for technology adoption have been rarely studied.[11] Notable exceptions include Atkin et al. (2015) who examine how temporary financial incentives can overcome workers resistance to adopt a new and more efficient technology to produce soccer balls. The authors find that initial slow-downs in productivity from learning a new production process inhibited the adoption of the technology in firms where workers where compensated for performance. The results show that a short run financial incentive large

9. See for example Basinga et al. (2011), Flores et al. (2013), Bonfrer et al. (2013), De Walque et al. (2015), De Walque et al. (2015), Gertler and Vermeersch (2012), Gertler et al. (2014), and Huillery and Seban (2014). Miller and Babiarz (2013) provide a review.

10. See Baker et al. (1988), Holmstrom and Milgrom (1991), Gibbons (1997), and Lazear (2000).

11. There is also work on temporary incentives in the form of sales and coupons to market products: e.g. Blattberg and Neslin (1990), Kirmani and Rao (2000), and Dupas (2014). Similarly, there is a literature on the effect of temporary incentives for individuals to develop better health habits such as exercise and quitting smoking– e.g. Volpp et al. (2008), Volpp et al. (2009), Charness and Gneezy (2009), John et al. (2011), Royer et al. (2012), Cawley and Price (2013), and Acland and Levy (2015).

enough to compensate workers for their short-term loss generated long run gains in productivity. In another related paper, Duflo et al. (2011) study the effect of providing short-term subsidies to purchase new more effective fertilizer by small farmers. The authors argue that even though new fertilizer is highly profitable, there might be important behavioral barriers and direct costs that inhibit their adoption. They show that small and temporary subsidies generated large increases in adoption, especially among impatient farmers. Finally, Bloom et al. (2012) show that management practices explain a great part of the differences in productivity among Indian firms in the textile industry. They show that providing managers with free short-term consulting on better management skills can create large gains in productivity in the long run.

The paper is organized as follows. Section 2.2 describes a simple model of technology adoption under fixed costs. Section 2.3 describes the intervention and the experimental design. Section 2.4 describes the data. Section 2.5 explains the identification strategy and estimation methods. Section 2.6 shows our main results on different outcomes and discusses the main mechanism to explain the effects we find as well as alternative explanations to our results. Section 2.7 and section 2.8 discusses spill over effects and effects on birth outcomes, respectively. Section 2.9 concludes.

2.2 Conceptual Framework

We develop a stylized model where clinics incur a fixed cost to change clinical practice patterns. We use this model to investigate the decision to adopt a new service into the set of practice patterns where the marginal return to the service is profitable but there is a fixed cost of adoption. We assume that patients are identical, that clinics provide the same services to all patients, and that demand is exogenously determined.

Objective Function: Clinics have a pay-off function $R = \pi + \alpha HN$, where π is profits, H is health of the representative patient, N is the number of patients, and $\alpha \geq 0$ is the

provider's intrinsic value of a unit of patient health.[12] When α takes on value 0, the clinic is purely extrinsically motivated and as α rises the clinic is willing to sacrifice more income for patient health. While we allow for both extrinsic and intrinsic motivation in the model, all of the results follow even with pure extrinsic motivation. Allowing for intrinsic motivation does not change the direction of the predictions just the magnitude.[13]

Health Production Function: Treatment technology, as defined by CPGs, involves two services, S_1 and S_2 where $S_i = 1$ if the clinic provides the service and 0 if not. If the clinic provides both services, then it is operating at the production possibilities frontier. The health production function for the representative patient is $H = \lambda_1 S_1 + \lambda_2 S_2 + \varepsilon$, where ε is a mean zero random shock.

Clinical Practice Patterns: Consider a clinic whose current clinical practice pattern is to provide S_1 to all patients. In this case, S_1 is the clinic's existing clinical practice pattern, and S_2 is an additional service that the clinic could choose to add to its practice routine. If the clinic wants to integrate the provision of S_2 into its practice pattern then it must incur an upfront fixed cost F. Fixed costs include designing, testing, and learning how to best incorporate the delivery of the service into existing practice patterns, retraining, purchase of complementary medical equipment, and reduced productivity during a period of adjustment.

Profits: Clinics are paid p_i for S_i and the marginal cost of providing S_i to a patient is c_i. Clinic's profits can then be expressed as:

$$\pi = \sum_{t=1}^{\infty} \beta^t [(p_1 - c_1) + (p_2 - c_2) S_2] N - F S_2 \tag{2.1}$$

where β is the clinic's discount rate. The discount rate may in part reflect present bias and psychological resistance to change; discounting future returns to an innovation at a

12. There is evidence to support intrinsic motivation as at least partially motivating medical care providers. See for example Leonard and Masatu (2010), Kolstad (2013), and Clemens and Gottlieb (2014).

13. Without some fixed costs of adjustment, both intrinsically and extrinsically motivated providers would still operate at the efficient frontier.

higher rate thereby lowering the present value of an innovation.

Adoption: The clinic adopts S_2 if

$$R(S_2 = 1) - R(S_2 = 0) \geq 0 \tag{2.2}$$

Substitution of 2.1 and 2.2 into the pay-off function and rearranging terms allows us to write the condition as:

$$\sum_{t=1}^{\infty} \beta^t [p_2 - c_2 + \alpha\lambda_2] N \geq F \tag{2.3}$$

Clinics are more likely to adopt S_2 if the profit margin from S_2 is higher, they have higher patient volumes, and they have lower discount rates. Clinics who are more intrinsically motivated (i.e. higher α) are also more likely to adopt and maybe even willing to lose money in order to adopt S_2, especially if S_2 is very productive (i.e. higher λ_2).

There are two cases under which the clinic will not adopt the new service. The first is when the value of the service is negative, i.e., $(p_2 - c_2 + \alpha\lambda_2) < 0$. In this case, the clinic will never adopt S_2, even if $F = 0$. The second is when the service is valuable, i.e., $(p_2 - c_2 + \alpha\lambda_2) \geq 0$, but the net present value of adopting a valuable service is less than the fixed costs of adoption, i.e., $\sum_{t=1}^{\infty} \beta^t [p_2 - c_2 + \alpha\lambda_2] N < F$.

Temporary Incentives: In the first case, where the value of S_2 is negative and $F = 0$, an increase of θ in p_2 such that $p_2 + \theta + \alpha\lambda_2 > c_2$ will lead the clinic to provide S_2 as long as the increase is active. Once p_2 returns to its original level, the clinic will stop providing S_2.

In the second case, where the value of S_2 is positive but $F > 0$, a temporary increase in p_2 can induce the clinic to offer S_2 permanently even after p_2 reverts to its original level. Consider an increase of θ in p_2 in period 1 that disappears in subsequent periods.[14] Without

14. The alternative is a lump sum payment that is vulnerable to the possibility of noncompliance and maybe difficult to verify. However, a temporary increase in p_2 requires the clinic to change routines and actually adopt S_2 in order to get paid. In this sense the temporary price increase also includes a commitment

loss of generality we can simplify the model to 2 periods with β as the discount rate. In this case, based on 2.3, the increase of θ in p_2 in period 1 necessary to induce the provider to adopt S_2 is:

$$\theta \geq \frac{F}{N} - (1+\beta)[p_2 - c_2 + \alpha\lambda_2] \tag{2.4}$$

The temporary incentive, θ, at minimum covers the remainder of the fixed cost of adjustment that is not paid by the discounted present value of the future stream of surplus generated from the provision of S_2. The incentive goes down with scale N, the profit margin $(p_2 - c_2)$, the extent to which clinics are extrinsically motivated times the marginal product of S_2 in the health production function $(\alpha\lambda_2)$, and the discount rate.

Cross-Price Effects: One concern voiced in the literature is that price increases for some services might lead to a reallocation of effort from other services that remain unchanged leading to negative cross-price effects. The implicit underlying model in these papers is an individual physician allocating time between activities with a time budget constraint. In our model of a medical care organization that can hire more staff, cross-price effects are generated based on the nature of economies of scope in either the health care production function or cost function. If both the production and cost functions are additively separable, then there are no cross-price effects. If the functions are not separable, then it is possible to have either negative or positive cross-price effects depending the nature of substitutability in the production and cost functions.

2.3 Context and Experimental Design

The field experiment was conducted by *Plan Nacer*, a public insurance program that began in 2005 to improve access to quality health care for otherwise uninsured pregnant women and

device and hence is ex ante preferable.

children less than 6 years old (Gertler et al., 2014). Like Medicaid in the U.S. and *Seguro Popular* in Mexico, the national *Plan Nacer* program transfers funds to local governments, in this case Provinces, who are then responsible for enrolling beneficiaries, organizing the provision of services, and paying medical care providers. An innovative feature of the Argentine program is that it uses financial incentives to ensure that beneficiaries receive high-quality care. Financing from the National level to Provinces is based for 60% on program enrollment and for 40% on performance.

Provinces then use those funds to pay public health care facilities on a fee-for-service basis for health care provided to program beneficiaries. The national government determines the content of the benefits package, which is uniform across provinces, while provincial governments set the price they will pay to providers for each service in that package. Revenues from *Plan Nacer* are on top of clinic budgets that cover salaries as well as medical and non-medical supplies and materials. In practice, *Plan Nacer* payments top up these budgets by 5 to 7%. Health facilities are free to choose how to use realized revenues within relatively broad guidelines, and in Missiones clinics can and do use 50% of the *Plan Nacer* payments to pay bonuses to clinic staff. In this sense, all services, including prenatal care visits, provided to Plan Nacer beneficiaries are in the interest of clinic staff as 50% of the payment is used to increase staff bonuses.

Plan Nacer scaled up by first recruiting and training clinics in the operations of its program, including fee structure, billing, and other rules. The program regularly retrains the clinics to keep them up to date on any changes and reinforce areas that are perceived to be weak. After clinics are enrolled, clinic community outreach staff identifies eligible women and children in order to enroll them into the program. Enrollment activities usually consist on door-to-door visits across a determined geographic area assigned to each clinic and defined by the Province.

Clinics can only provide services to the population within their area and enrolled ben-

eficiaries can only obtain care from their assigned clinic. Outreach staff regularly contact beneficiaries to encourage them to take advantage of program benefits. *Plan Nacer* reimburses clinics for all outreach activities to the beneficiary population at a rate higher than the clinic's cost of outreach.

The field experiment was conducted with primary health care clinics in the Province of Misiones, one of the poorest in the country and with high rates of maternal and child mortality. In Misiones, each clinic is allowed to use up to 50% of revenue from *Plan Nacer* fees to pay bonuses to facility personnel at the discretion of the facility director. The rollout of *Plan Nacer* in Misiones was completed in 2008 long before the pilot study. As such, both providers and beneficiaries were knowledgeable of the operation of Plan Nacer before the experiment began.

The experimental intervention was designed to encourage early initiation of prenatal care for *Plan Nacer* beneficiaries, thereby aligning the incentives in *Plan Nacer* with official Argentine clinical practice guidelines, medical school training, and international scientific evidence. Before the experiment, only one-third of Plan Nacer beneficiaries were initiating care in the first trimester (MINSAL, 2009a,b). The experiment randomized temporary financial incentives to primary health care clinics in which treatment clinics were paid a 200% premium for early initiation of prenatal care, i.e. before week 13 of pregnancy.

Table 2.1 presents the payment schedule for the periods before, during and after the intervention. Prior to the intervention period, the province paid facilities 40 ARS for each prenatal visit regardless of when it occurred or whether it was the first or a subsequent visit.[15] At this initial price prenatal care visits were profitable as 50% of this fee was used to increase staff bonuses. During the intervention period the fee was increased to 120 ARS for 1st visits that occurred before week 13 but remained at 40 ARS for subsequent visits. Every other component of the Plan Nacer program remained the same. After that, the

15. The exchange rate for 1 ARS was around 0.25 USD between 2009 through 2011.

intervention period fees reverted to the original payment of 40 ARS for all visits. The modification amounted to a 3-fold increase in the fee for 1st visits before week 13. The modified fee structure was implemented for 8 months - from May 2010 to December 2010.

Table 2.1: Payments for First Prenatal Visit

Time Period	Dates		Payment for 1st Prenatal Visit	
	Begin	End	Preg. Week <13	Preg. Week ≥13
Pre-Intervention	January 2009	April 2010	$ 40 ARS	$ 40 ARS
Intervention	May-10	December 2010	$ 120 ARS	$ 40 ARS
Post Intervention	January 2011	December 2012	$ 40 ARS	$ 40 ARS

Notes: Source from MINSAL (2009b).

Facilities selected to receive the modified fee structure were invited to participate and notified of the time-limited implementation on April 14, 2010. Facility directors were required to sign a formal modification of their existing contract with Plan Nacer in order to receive the modified fee structure.

The study design included 37 clinics out of 262 primary care facilities of the province, of which 18 were randomly assigned to the treatment group and were offered the modified fee schedule. The other 19 formed the control group. Table 2.2 shows that compliance with treatment assignment was not perfect: out of 18 facilities assigned to the treatment group, 14 were actually treated as three refused to sign the agreement and a fourth closed before the intervention started. In addition, one of the facilities originally assigned to the control group was mistakenly offered the treatment and agreed to the modified fee structure. In the end, there were 36 facilities in the study excluding the one that closed.

2.4 Data

The Province of Misiones maintains a well-developed and long-established automated medical record information system managed by the provincial authorities. Personnel at public

Table 2.2: Clinic Assignment and Compliance Status

Randomized	Compliance		
	Yes	No	Total
Yes	14	4	18
No	1	18	19
Total	15	22	37

Notes: Authors' own elaboration.

primary health clinics and hospitals digitize a record of each service provided to each patient. The data are of unusually high quality in that key outcomes such as dates of visits, services delivered, and birth weight are recorded at the time each service is provided; therefore we do not need to rely on maternal recall of these variables usually collected by surveys long after the visit. The data used in the analysis are extracted from individual clinical records and contain information on the universe of patients for the 36 clinics in the study. The records also include the individual's national identity number, which is used to link the individual clinic medical records from primary health facilities with the registry of health insurance coverage, the registry of Plan Nacer beneficiaries, and hospital medical records. In all, 97% of the primary clinic medical records were merged with the data on insurance status and program beneficiary status. In addition, 75% of these were successfully merged with medical records data from hospitals. Therefore, each observation in our sample corresponds to a unique pregnancy by women who initiated their prenatal care in one of the primary care clinics of the sample.

2.4.1 Analysis Sample

The timeline of the study and the availability of data is divided into 4 different sub-periods: (i) a 16-months pre-intervention period from January 2009 to April 2010, (ii) an 8-month intervention period from May 2010 to December 2010, (iii) a 15-month "post-intervention

period I" from January 2011 to March 2012 and (iv) a 9-month "post-intervention period II" from April 2012 to December 2012. Prenatal care data was consistently collected for the first 3 periods from January 2009 through March 2012. Starting in April 2012, however, Misiones adopted a new information system and as a result data from post-intervention period II cannot easily be compared to data from the earlier periods. In particular, the new system changed the codes used to classify the reason for visits in order to facilitate billing. If in the first visit the attending physician requested an ultrasound to confirm a pregnancy, this first visit was labeled as a "care visit" while the subsequent (second) visit, was labeled as the first prenatal visit, if indeed the ultrasound confirmed the pregnancy. On average, this would lead to a reduction in the share of women who had a visit labeled as "first prenatal visit" before week 13 and an increase in the weeks pregnant at the time of this visit. If the new coding system affected the treatment and control groups in the same way, the differences between the treatment and control groups would still capture the impact of the incentives, albeit possibly with some measurement error. Therefore, we analyze the data from post-intervention period II separately, and interpret the results with caution.

The analysis sample includes pregnant women who were beneficiaries of Plan Nacer at the time of the first prenatal visit.[16] While information on prenatal care utilization is available for the full sample period, information related to birth outcomes is only available for women who gave birth in a public hospital through 2011, i.e. women that became pregnant before May 2011.

2.4.2 Measurement of Weeks Pregnant at First Prenatal Visit

Each observation in our sample corresponds to a different pregnancy that initiated prenatal care in one of the clinics included in the experiment. For each pregnancy we observe the

16. We excluded non-beneficiaries because most of them have private health insurance and as such are likely to receive some of care and deliver at private facilities. Since we do not have data from private facilities, the outcomes of most of these observations are censored.

date of the first prenatal visit and the date of the last menstruation period as recorded by the physician. We construct the number of weeks of pregnancy at the time of the first prenatal visit as the difference between the date of the first visit and the last menstrual date (LMD). The LMD is routinely collected at the time of the visit to calculate the estimated date of delivery (EDD) and both are routinely recorded in the patient's medical record at the clinic.[17]

One potential problem is that medical personnel in treatment facilities might misreport the date of the first visit as occurring before week 13 so that they could bill it to the program at a higher amount. We think this is unlikely for the following reasons. First, the week of pregnancy at the first visit is constructed from the date of the first prenatal visit and the LMD, both of which along with the EDD are recorded in real time in the medical record. In order to falsely report that a first visit occurred in the first 12 weeks, the provider would have to alter the date of the first visit relative to either the LMD or the EDD in the medical record. This would require some effort if done in real time and would be noticeable by auditors if altered ex post. Second, Plan Nacer uses external auditors to verify the accuracy of clinic billing. The auditors compare the detailed clinical records to the billing requests to find inconsistencies that could turn into substantial financial penalties for the provinces. Third, while there may have been an incentive to misreport during the intervention period, there was no financial return to misreporting in the post-intervention period once the incentives were removed. It also was unlikely that it was worth the clinics' time to set up elaborate procedures for falsifying records when they knew the incentives were only in place for 8 months. Finally, clinical records are legal documents in Argentina and practitioners could lose their medical license if caught systematically misreporting for financial gains.

To corroborate our belief that false reporting of records is unlikely, we empirically test whether there is any evidence of systematic misreporting using data from an alternative

17. For 10% of the sample LDM was not recorded. For those cases, we use the EDD to recover the LMD.

source. Specifically, we use gestational age at birth measured by physical examination obtained from hospital records to construct a second estimate of the LMD and weeks pregnant at the time of the first prenatal visit. The hospital personnel that attend the birth do not have any incentive to misreport hospital records. We then compare the estimated week of first visit based on gestational age at birth to the week of first visit reported by the health facilities. The results do not show any evidence of systematic misreporting due to incentives. Appendix A provides a detailed discussion of the analysis and results.

We also explore whether there is any manipulation of the data at the threshold of the 13th week of pregnancy. Appendix A Figure A.2 shows that there is no discontinuity at this threshold using the test proposed by McCrary (2008) for manipulation at the threshold in studies that use Regression Discontinuity as their research design.

2.4.3 Descriptive Statistics and Baseline Balance

Table 2.3 reports the descriptive statistics for the key outcomes of interest and demographic characteristics at baseline, i.e. in the 16-month pre-intervention period (Jan 2009 – April 2010). Outcomes are balanced at baseline in that there are no statistically significant differences in the means of variables between the treatment and control groups. On average women had their first prenatal visit about 17.5 weeks into their pregnancy with about one-third of women having that visit before week 13. Women completed about 4.7 prenatal visits over the course of their pregnancy and more than 80% of them received a tetanus vaccine. Newborns weighed approximately 3,300 grams on average, while about 6% of them were born with low birth weight (i.e. less than 2,500 grams), and slightly more than 9% of births were born prematurely.

Table 2.3: Baseline Descriptive Statistics

	Assigned Treatment Group		Assigned Control Group		p-Value for test of equality of means	
	Mean (s.d.)	N	Mean (s.d.)	N	Large sample	Wild Boot-Strapped
Weeks Pregnant at 1st Prenatal Visit	17.5	743	17.6	497	0.89	0.84
	(7.48)		(7.74)			
1st Visit before Week 13 of Pregnancy	0.35	743	0.33	497	0.57	0.56
	(0.48)		(0.47)			
Tetanus Vaccine During Prenatal Visit	0.80	743	0.84	497	0.34	0.41
	(0.40)		(0.37)			
Number of Prenatal Visits	4.68	743	4.28	497	0.39	0.45
	(2.94)		(2.77)			
Birth Weight (grams)	3.328	552	3.291	379	0.36	0.37
	(519)		(558)			
Low Birth Weight ($<$ 2500 grams)	0.06	552	0.06	379	0.96	0.98
	(0.23)		(0.23)			
Premature (gestational age $<$ 37 weeks)	0.09	319	0.10	249	0.83	0.82
	(0.29)		(0.30)			

Notes: This table presents means and standard deviations in parentheses for the treatment and control groups during the 16-month pre-intervention period from January 2009 through April 2010. P-values for tests equality of treatment and control groups means are presented in the last 2 columns. We present both the p-value computed for large samples and a Wild bootstrapped p-value that is robust in samples with small numbers of clusters (Cameron et al., 2008). Our Wild bootstrap procedure assigns symmetric weights and equal probability after re-sampling residuals (Davidson and Flachaire, 2008) and uses 999 replications.

2.5 Identification and Estimation

We estimate both the intent-to-treat (ITT) and local average treatment (LATE) effects of the incentives on outcomes (Guido W. Imbens, 1994). The ITT is the effect of assigning a clinic to treatment on outcomes, regardless of compliance. The LATE is the effect of a clinic actually receiving the incentives. In both cases, the treatment effect is identified off the variation induced by the randomized assignment status. In the discussion of results in

the next section, we report the LATE estimates.[18]

The ITT estimate compares the mean outcome of the group assigned to treatment to the mean outcome of the group assigned to control and is estimated by regressing the outcome against an indicator of whether the clinic was assigned to treatment using the following specification

$$y_{ijt} = \alpha_t + \beta_t D_j + \varepsilon_{ijt} \tag{2.5}$$

where y_{ijt} is the outcome of pregnancy i receiving care in clinic j in period t, T_j is a dummy variable taking on the value 1 if the clinic was assigned to the treatment group and 0 otherwise, and ε_{ijt} is a zero mean random error. Notice that parameters are allowed to vary by period. We work with four different periods of analysis: an 18-month pre-intervention period, an 8-month intervention period, a 15-month post intervention period I, and an 8-month post intervention period II. We estimate separate models for each of these periods. In the LATE model we replace T_j the "assigned to treatment" variable with an indicator of being actually treated and use the clinic's randomized assignment status as an instrumental variable for actual treatment.

Our sample is clustered within 36 health clinics since the random assignment of treatment occurred at the clinic level. As such, there may be intra-cluster correlation that must be considered for statistical inference. Standard methods of correcting standard errors rely on large sample theory both in the number of observations and in the number of clusters. Given the small number of clusters in our sample, we instead use randomization inference methods that are robust to randomized assignment of treatment among a small number of clusters. Specifically, we use the Wild bootstrap to generate p-values for hypothesis testing in ITT models (Cameron et al., 2008) and an analogous method for hypothesis testing in the LATE

18. The ITT results are almost identical to the LATE estimates, which is expected given the relatively high compliance rates to the original assignment. The ITT results are presented in Appendix C.

models (Gelbach et al., 2009). Our Wild bootstrap procedure assigns symmetric weights and equal probability after re-sampling residuals, and uses 999 replications (Davidson and Flachaire, 2008).

2.6 Timing of First Prenatal Visit

In this section we report the results of analyses of the effects of the temporary incentives on the timing of the first prenatal visit and mechanisms by which clinics achieved those results.

2.6.1 Densities

Figure 2.2 compares the densities of weeks pregnant at the time of the first prenatal visit for the clinics assigned to the treatment and control groups and reports p-values for Kolmogorov-Smirnov tests of equality of the distributions. Panel A shows that there is no difference between the densities of the treatment and control groups in the pre-intervention period. Panel B shows that the treatment group density is to the left of the control group density during the intervention period. Finally, Panel C and D show that the treatment group density is placed to the left of the control group density during post-intervention periods I and II. Kolmogorov-Smirnov tests for equality of the distributions cannot be rejected for the pre-intervention analysis, but are rejected for the intervention and both post-intervention periods with p-values of 0.031, 0.004, and 0.009 respectively. These results imply that the temporary incentives led to earlier initiation of care in the treatment group compared to the control group in the intervention period and that these higher levels of care persisted for at least for 24 months and more after the higher fees were removed.

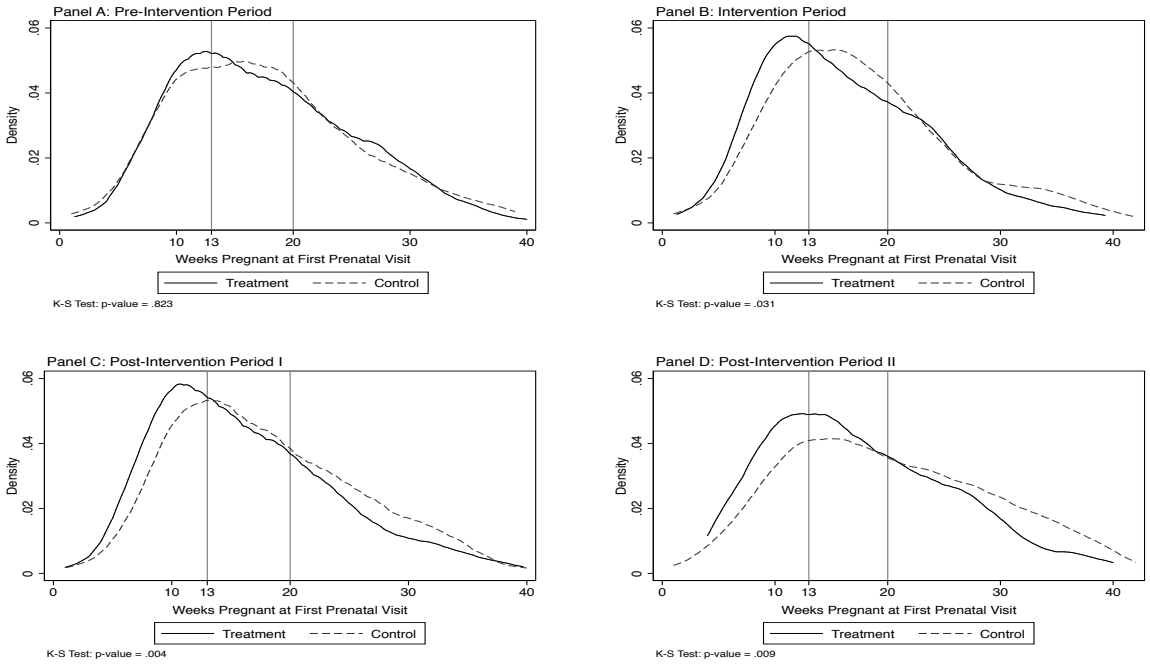

Figure 2.2: Densities of Weeks Pregnant at First Prenatal Visit

Notes: Densities estimated using an Epanechnikov kernel with optimal bandwidth. P-vales of Kolmogorov-Smirnov tests of equality of distributions between groups reported below figure. The two vertical lines indicate weeks 13 and 20 of pregnancy. Source: Authors' own elaboration based on data from the provincial medical record information system.

2.6.2 Short Run Effects

Table 2.4 reports the estimates of the effects of the temporary fees on the early initiation of care. Panel A reports the results for weeks pregnant at the time of the first prenatal visit and Panel B reports the results for whether the first visit occurred before week 13. The first column reports the results for the intervention period and the second and third columns report the results for the post-intervention periods. During the intervention period, on average, women in the treatment group had their 1st visit about 1.5 weeks earlier in their pregnancy than women in the control group. The share of women in the treatment group who had their 1st visit before week 13 is 11 percentage points higher than the control group; approximately 35% higher than the control group. Both estimates are significantly different

from zero at conventional p-values.

Table 2.4: Effects of Temporary Incentives on Timing of First Prenatal Visit

	(1)	(2)	(3)
	Intervention Period	Post-Intervention Period I	Post-Intervention Period II
A. Weeks Pregnant at 1st Prenatal Visit			
Treatment	-1.47**	-1.63**	-2.47**
	(0.71)	(0.75)	(1.02)
Large Sample p-value	0.04	0.03	0.02
Wild Bootstrapped p-value	0.08	0.03	0.03
Control Group Mean	17.80	17.90	20.10
Sample Size	769	1.296	710
B. First Prenatal Visit Before Week 13 of Pregnancy			
Treatment	0.11**	0.08**	0.08**
	(0.04)	(0.04)	(0.04)
Large Sample p-value	0.01	0.02	0.04
Wild Bootstrapped p-value	0.03	0.05	0.06
Control Group Mean	0.31	0.34	0.27
Sample Size	769	1.296	710

Notes: This table reports LATE estimates of the treatment effect estimated from 2SLS regressions of the dependent variable on actual treatment status instrumented with clinic treatment assignment type. The p-values are for tests of the null that the difference is equal to zero. We present both the p-value computed for large samples and a Wild bootstrapped p-value that is robust in samples with small numbers of clusters (Cameron et al., 2008). Our Wild bootstrap procedure assigns symmetric weights and equal probability after re-sampling residuals (Davidson and Flachaire, 2008) and uses 999 replications. Standard errors are in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0$

2.6.3 Long Run Effects

Our model in Section 2.2 provided clear predictions about provider behavior once temporary incentives disappear: i.e. if the fee increase is enough to overcome the fixed costs of adapting a new practice, clinics should maintain higher levels of prenatal care after incentives are removed. Column 2 of Table 2.4 reports the estimated impact of the temporary fee increase on early initiation of care in the 15-month period after the fees were removed. On average,

pregnant women in the treatment group started their care 1.6 weeks earlier than those in the control group. The difference between the treatment and control groups in the share of women who had their 1st visit before week 13 was 8 percentage points. Both estimates are statistically different from zero at conventional levels. Further, we cannot reject the null hypothesis that the impact is different in the intervention and post-intervention periods.

While there is no significant difference between the effect during the intervention and the post-intervention periods, one concern may be that the effect of treatment slowly trended towards zero after the incentives ended. To explore this hypothesis, we plot the mean number of weeks pregnant at the time of first prenatal visit for treatment and control groups, before, during and after the intervention (Figure 2.3).[19] We split the pre-intervention period into two sub-periods of 6-months each and the post-intervention period into 3 sub-periods: the first two are 6 months and the third is 3 months. The treatment effect is the difference between the two lines. While the treatment and control groups have similar trends before the intervention, the treatment group appears to receive earlier care during the intervention, and the change persists after the end of the intervention. Notice that there is little if any fall off over the post-intervention period. Rather, the treatment effects remain fairly constant over the 15 -month post-intervention period I. Figure 2.4 depicts the same relationship for the share of women who receive care before week 13 of pregnancy.[20] Again, the effects of the intervention appear to continue at a steady rate after it is discontinued.

2.6.4 Longer Run Effects

The period of analysis in our main results is restricted to January of 2009 to March of 2012. Recall that starting in April 2012, the visit coding system changed. Hence starting in April 2012 what is reported as first visit in the data is actually a mix of first and second

19. As discussed above, the information from post-intervention period II (April-December 2012) uses a different metric and is therefore not included in this figure.

20. Ibidem.

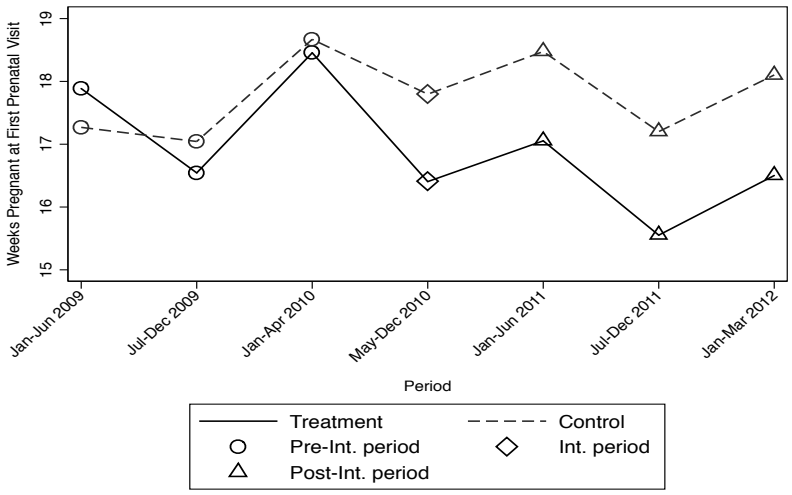

Figure 2.3: Mean Number of Weeks Pregnant at First Prenatal Visit

Notes: The first two points (circles) are means for 6-month periods prior to the intervention period. The third point (diamond) corresponds to the 8-month intervention period. The fourth and fifth points (triangles) correspond to 6-months periods after the intervention period, while the last point (triangle) is for a 3-month period.

visits. As a result the average of weeks pregnant at the first visit increases and the share of pregnant women whose first visit was before week 13 falls relative to previous periods. Column 3 in Table 2.4 shows the results for this last period. The mean average of weeks pregnant at the time of the first visit for the control group is substantially higher for this period than for previous periods and the mean share that had their first visit before week 13 is substantially lower, suggesting that there is measurement error in our main outcome in this period. However, this difference in coding should have a similar effect in treatment and control clinics given the randomized assignment of the treatment. Therefore the difference between treatment and control clinics should cancel out the measurement error and provide us with unbiased estimates of the impact.

The results in Table 2.4 show a statistically significant reduction in the number of weeks pregnant at the time of the first visit and a statistically significant increase in the share of pregnant women who had their first visit before week 13. These results suggest that the

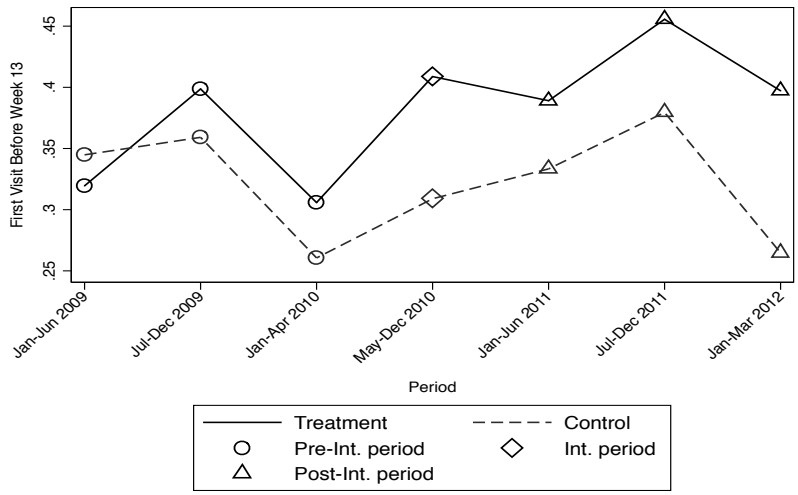

Figure 2.4: Proportion of Mothers with First Prenatal Visit before Week 13 of Pregnancy

Notes: The first two points (circles) are means for 6-month periods prior to the intervention period. The third point (diamond) corresponds to the 8-month intervention period. The fourth and fifth points (triangles) correspond to 6-months periods after the intervention period, while the last point (triangle) is for a 3-month period.

improved productivity persisted at least 24 months after the fees were removed.

2.6.5 Robustness

We implement three robustness checks. First, the main sample may include pregnancies that start in one period and end in another, which could cloud the effect of the incentives on timing of the first visit. For example, a woman who is 6 months pregnant and had not had a prenatal visit when the intervention starts and subsequently receives her first prenatal checkup during the intervention, would be counted as a third trimester first visit during the intervention period, even though the intervention cannot affect whether she receives prenatal care before week 13. Hence, we re-estimate the models on a restricted sample where women are no more than one month pregnant in the first month of the period and no less than 3-months pregnant in the last month of the period. The results, reported in Panels B of Appendix B Tables B.1 and B.2, are very close in magnitude and statistical significance to

the main results in Table 2.4.

Second, even though there were no statistical differences in baseline means, it is possible that randomization was not able to fully balance the treatment and control groups on unobservable characteristics given the small number of clinics. In order to test for this possibility, we estimated the models using difference-in-differences with clinic and month fixed effects. The results, reported in Panels C of Appendix B Tables B.1 and B.2, are very close in magnitude and statistical significance to the main results in Table 2.4.

Finally, in studies involving a small sample of clusters there is a concern that a few outliers may drive the average effect found in the previous sections. We explore this possibility in two ways. First, we re-estimate the models by dropping all the observations from one clinic one at a time. This produces 36 different estimated treatment effects, which we picture in appendix Figures B.1 and B.2 for the outcomes of weeks pregnant at the time of the first prenatal visit and for the probability that the first visit occurred before week 13, respectively. The results are sorted along the x-axis from the lowest to the highest estimated effect, while the dashed blue line is the intent-to-treat effect calculated by pooling the intervention and the first post intervention period. The solid black line represents a zero treatment effect. The vertical lines are 95% confidence intervals constructed using standard errors obtained from the Wild bootstrap procedure. Notice that there is almost no difference in any of the estimates implying no one clinic drives the estimated effects in Table 2.4.

Second, we estimate clinic-specific treatment effects whereby we compare each treated clinic individually to the control clinics as a group. Appendix Figures B.3 and B.4 plot these individual clinic treatment effects for the outcomes of weeks pregnant at the time of the first prenatal visit and for the probability of that the first visit occurred before week 13, respectively. The results are again sorted along the x-axis from the lowest to the highest estimated effect, while the dashed blue line is the intent-to-treat effect calculated by pooling the intervention and the first post intervention period and the solid black line represents a

zero treatment effect. The figures show that the hypothesis of no treatment effect is rejected for 11 out of 17 clinics in Figure B.3 and 12 out of 17 clinics in Figure B.4. In addition, the treatment effects have the expected sign in 15 out 17 clinics in Figure B.3 and 14 out of 17 clinics in Figure B.4. This provides evidence that our results are not driven by a few large-effect clinics.

2.6.6 Mechanisms

In order to better understand how clinics were able to achieve such large increases in the share of women who initiated prenatal care before week 13, we conducted a series of in-depth interviews with medical professionals in 5 of the 14 actually treated clinics. In this section we first report what we learned from these interviews. In summary, all of the clinics reported developing a new set of community outreach activities designed to identify Plan Nacer beneficiary women early in their pregnancies and reach out to them to encourage early initiation of prenatal care. The design and installation of these outreach activities into clinic routines involved nontrivial fixed costs and the delivery of those services created new variable costs. We then use the whole sample to analyze the impact of the temporary incentives on community outreach activities.

Developing New Outreach Strategies.– All of the clinics reported organizing a team meeting with the staff at the beginning of the intervention in order to discuss and brain storm strategies to respond to the new incentive scheme. They developed innovative data driven strategies to identify women who were likely to be pregnant. The clinics then typically sent staff to inquire about last menstruation date and offer an instant-read pregnancy test to those women whose menstruation was overdue. If pregnant, they then encouraged the expectant mothers to start prenatal care quickly.

Much of this involved experimenting with different strategies until they found what worked best. For instance, health workers started to monitor women who used birth control

pills and prioritize home visits to women who were late in picking up their pill refills.[21] Second, clinics targeted mothers who already have children, as they are less likely to initiate their prenatal visits early in a new pregnancy, by meeting them at free milk distribution centers.[22] Third, health workers noted that adolescents are less willing to reveal a pregnancy in the presence of their parents. Clinics therefore changed the timing of home visits so as to increase the chance of finding adolescents by themselves. Clinics' work schedules were also modified so as to ensure predictable availability of a gynecologist on certain days of the week so health workers could better schedule patient appointments. Other clinics started keeping track of visits to "at risk" patient and map clinic catchment areas with corresponding (potential) pregnancies so as to more efficiently organize home visit routes.

Implementation and Cost of New Outreach Strategies.– Clinics used Community Health Workers (CHWs) to implement these new activities.[23] Normally, CHWs carry out community outreach activities including promotion of preventive health, follow-up of patients in treatment including pregnant women, follow-up of immunization status of children, health data management, early detection of malnutrition in children, among others as well as periodically updating the roster of residents in the clinic's catchment areas. Since its rollout in 2005, Plan Nacer has reimbursed clinics for outreach activities at a profitable rate.[24] CHWs work under temporary contracts of variable length with the facilities and are not part of the formal civil service subject to more rigid labor laws. As such, clinics can easily and quickly

21. Birth control pills are dispensed free of charge by each health facility. Women cannot collect more than a months supply at any one time and must return each month for refill. The pharmacy unit keeps records of birth control pill collections.

22. Plan Nacer beneficiaries with young children are eligible for free milk weekly and mothers collect the milk at distribution centers.

23. The Ministry of Health created CHW as a job category in 2005 as part of a 3-year associates degree program in Primary Health Sector Management from the Ministry of Health. CHWs have classes at least 4 hours per week and are required to work at least 21 hours a week as interns in a local clinic or hospital. The interns are paid an hourly stipend that is less than the minimum wage.

24. From administrative records we the average cost of outreach activities to 1 USD as CHWs are paid 2 USD per hours and complete on average 2 outreach activities per hour. Plan Nacer pays 2.5 USD for outreach activities to pregnant women, so that each outreach activity generates a profit of 1.5 USD.

expand and contract the amount of CHW labor they employ. During the intervention, clinics reported expanding CHW activities by increasing the hours of existing CHWs as opposed to hiring new CHWs and paid incentive bonuses to CHWs for getting pregnant women into prenatal care.[25]

Impact of Temporary Incentives on Outreach Activities.– We are able to substantiate the claims of increased CHW outreach activities using clinic administrative data for the whole sample.[26] Figure 2.5 displays the average and median number of CHW outreach activities that resulted in maternal care visits for the pre-intervention, intervention, and post-intervention I periods.[27] The results show that there is no difference in outreach activities between treatment and control clinics in the pre-intervention period. In the intervention period the treatment group had substantially more activities than the control group, and this difference continued through the post-intervention period. We use these data to estimate the differences in the logarithm of number of activities between the treatment and control groups using the same methods in Table 2.4. The results show no differences in activities in the pre-intervention period, and positive and statistically significant higher levels of activities in the treatment clinics in both the intervention and post-intervention periods (Table 2.5). Outreach activities doubled in the treatment clinics relative to the controls in both the intervention and post intervention periods suggesting that the temporary incentives significantly raised CHW outreach activities to a level that persisted at least 15 months after the temporary incentives were removed.

25. Until 2013 health facilities participating in Plan Nacer in Misiones was able to use up to 50% of their of Plan Nacer funds to pay bonuses to health professionals. The bonuses could be assigned to any person working at the health facility, including CHWs

26. Plan Nacer finances clinic outreach activities on a fee-for-service basis and employs an external independent auditor to audit clinic activity reports. Treatment and comparison clinics were paid the same fee for these activities before, during and after the experiment.

27. The medians are better measures of central tendency as the densities of both activities are asymmetric heavily skewed to the right.

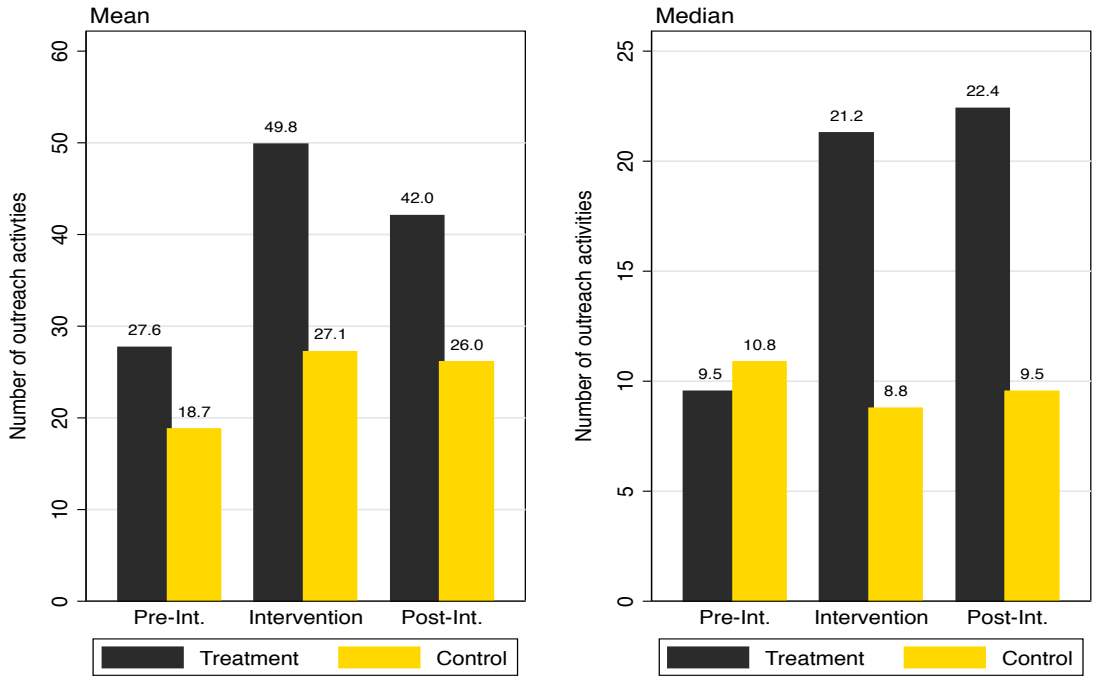

Figure 2.5: Number of Clinic Outreach Activities

Notes: The bars report the mean and median number of outreach activities that resulted in actual maternal-child service at the clinic, per trimester for the pre-intervention period (January 2009-April 2010), the intervention period (May-December 2010), and post-intervention period I (January 2011-March 2012).

2.6.7 Salience and Importance of Early Initiation of Prenatal Care

A key aspect of the argument that fixed costs of adoption inhibited clinics from adopting services to increase the early initiation of prenatal care is that clinics valued early initiation enough to have had adopted these services without the fixed costs. It is possible, however, that fixed costs of adoption were not inhibiting adoption, but rather clinics did not sufficiently value early initiation of care enough to invest in these services without the increased fees. The temporary incentives might have just made early initiation of care more salient and thereby increased the importance of early initiation of care in the staff's minds so that it

Table 2.5: Effects of Temporary Incentives on Log Number of Outreach Activities

	(1)	(2)
	Intervention Period	Post-Intervention Period I
Treatment	0.47**	0.56**
	(0.23)	(0.22)
Large Sample p-value	0.04	0.01
Wild Bootstrapped p-value	0.04	0.02
Log (Control Group Mean)	1.93	1.93
Sample Size	324	324

Notes: This table reports LATE estimates of the treatment effect estimated from 2SLS regressions of the dependent variable on actual treatment status instrumented with clinic treatment assignment type. The p-values are for tests of the null that the difference is equal to zero. We present both the p-value computed for large samples and a Wild bootstrapped p-value that is robust in samples with small numbers of clusters (Cameron et al., 2008). Our Wild bootstrap procedure assigns symmetric weights and equal probability after re-sampling residuals (Davidson and Flachaire, 2008) and uses 999 replications. Column (1) reports the results for the sample observed in an 8-month intervention period (May 2010 - December 2010). Column (2) reports the results for the sample observed in the 15-month period following the end of the intervention (January 2011 - March 2012).). Standard errors are in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0$

became a higher priority for action.[28] Kahneman (2012, pp. 8) states that "... frequently mentioned topics populate the mind..." more than others and "... people tend to assess the relative importance of issues by the ease with which they are retrieved from memory". As such, salience "... is enhanced by mere mention of an event" Kahneman (2012, pp. 331). If incomplete or non-adoption of a task is a matter of salience as opposed to fixed costs of adoption then the observed treatment effects may be explained by the fact that temporary incentives help to overcome this type of psychological resistance to change.

While we do not have information on the salience of early initiation of care before the

28. Taylor and Thompson (1982) define salience as, "... the phenomenon that when one's attention is differentially directed to one portion of the environment rather than to others, the information contained in that portion will receive disproportionate weighting in subsequent judgments". See Bordalo et al. (2012, 2013) for a more recent discussion of salience and choice theory. See De Mel et al. (2013), and Karlan et al. (2016) for empirical analysis of salience effects through informational reminders.

experiment, we are able to explore whether the temporary fee increase made early initiation of care more important in the minds of the clinic staff after the end of the experiment. To do so we administered a survey to the chief medical officer of each clinic about the absolute and relative importance of seven different prenatal care procedures including initiating prenatal care prior to week 13 of pregnancy (see Appendix D).

Figure 2.6 compares the absolute score and relative ranking of the procedures in terms of importance for prenatal care. The absolute scores ranges from 0 to 5, with 5 being the highest while the relative ranking sorts the seven practices from 1 to 7, with 1 being the highest ranking. Our outcomes of interest are the absolute score and relative ranking assigned to early initiation of prenatal care. Panel A in Figure 2.6 shows that the absolute score assigned by medical directors to early prenatal care is on average 4.8 in the treatment group and 4.7 in the control group. Panel B in Figure 2.6 shows that on average the relative ranking for this practice is also similar between the two groups, 2.0 for the treatment group and 1.9 for the control group. Moreover, these differences are not statistically significant at conventional levels (see Appendix D). These results suggest that the early initiation of prenatal care is of very high absolute and relative importance, and that the temporary fees did not have an effect on either the absolute or relative importance of this practice.

2.6.8 Alternative Explanations

Substitution.– One alternative explanation for the short-term treatment effects is that the incentives are causing treatment clinics to try to attract pregnant women who otherwise would have used other clinics. This is unlikely to be true as beneficiary women are assigned to specific clinics when enrolled in Plan Nacer and cannot simply go to another clinic to receive care. Moreover, clinics and their CHWs have specific geographic areas assigned and do not conduct outreach activities outside of those areas. Finally, the number of patients per month and the share that initiate care before week 13 are the same in the pre- and

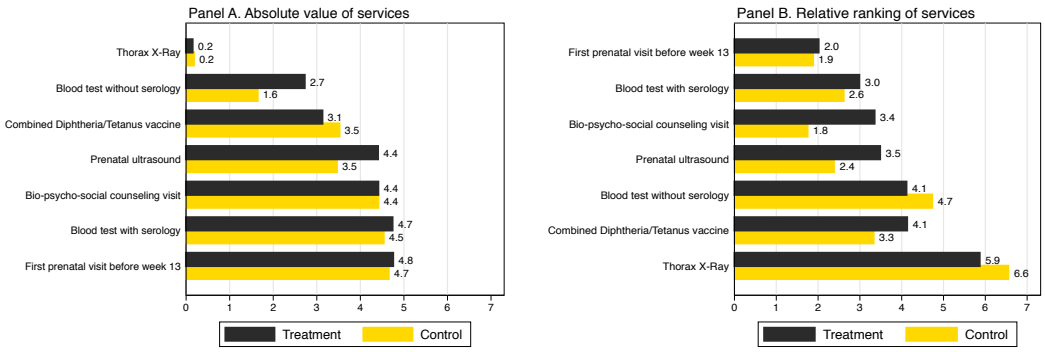

Figure 2.6: Importance of Prenatal Care Services

Notes: Panel A and Panel B report the average of the absolute score and relative ranking, respectively, that measures the importance given by clinics to seven different prenatal care procedures including initiating prenatal care prior to week 13 of pregnancy (Appendix D). The absolute scores range from 1 to 5, with 5 being the highest score in terms of importance.

post-intervention periods for control clinics, and the average monthly number of patients is also the same in the pre- and post-intervention periods for the treatment clinics.

Information Spillovers.– Another alternative explanation for long-run results is that after the temporary incentives ended, women who were pregnant during the intervention periods passed the message of the importance of early initiation of care onto other beneficiary women who became pregnant during the post-intervention period. Hence, the persistence of the effect of the incentives might be caused by an informational spillover. However, the higher amount of community outreach activities in treatment clinics, the mechanism used to generate higher early initiation of care, continued into the post-experimental period at the same level as in the intervention period. Hence, if there were information spillovers in the post-intervention period, then one would expect to see higher treatment effects in the post-intervention period than in the intervention period.

Labor Contract Frictions.– An additional candidate explanation for not reducing outreach activities after the temporary fees disappeared was not that clinics valued early initiation of prenatal care without the fee increase, but rather that CHW employment contracts were

sticky so it was hard and costly to reduce CHWs. This is unlikely to be true for a number of reasons. First, continuing to provide unproductive outreach services was costly and clinics could have reassigned CHWs to other tasks or reduced their use. Second, most of the CHW expansion was through increasing CHW hours and not hiring new CHWs so it should have been easy to reduce hours to pre-intervention levels. Third, any new CHW hired could have been easily dismissed. CHWs are temporary interns employed on part time contracts. Fourth, since clinics knew that the temporary fees only lasted eight months when the program started, they would not have hired new CHWs on contracts for longer than that period. Contracts would then have had to be renewed or new CHWs hired in order to continue the higher level of outreach activities. Finally, even if contracts were sticky clinics should have been able to reduce some of the outreach activities in the follow-up period but we see no reduction.

Career incentives.– It is also possible that even if clinic directors did not value early initiation of prenatal care, they did not reduce outreach activities in the post intervention period because they were worried that these reductions would harm their careers (Ashraf et al., 2014). This is highly unlikely for a number of reasons. First, the Government regularly monitored clinic performance on a large number of indicators, none of which was early initiation of prenatal care. With career incentives at play money spent on outreach activities could have been better used on services for which the clinic was explicitly accountable. Second, the intervention only changed the fees for a short 8-month period of time and not a permanent change that might also signal a long-term change in priorities that might have shifted beliefs of clinical directors and staff about the importance of early prenatal care. Third, there was no accompanying information explaining the reason for the temporary fee change or that clinics would be assessed on this indicator.

2.7 Cross-Price Effects

While the modified fee schedule was designed to affect the timing of the first prenatal visit, providers to reduce effort supplied to other services, resulting in a lower provision of such services to patients. We test for this by estimating the effect of the incentives on the probability of pregnant women having a valid tetanus vaccine, and the number of prenatal visits. The results presented in Table 2.6 report no evidence of cross-price effects, positive or negative, in either the intervention period or in post-intervention period I. In fact, the levels of these services appear to be constant over time. While the concern about crowding-out is typically for a context of individual providers facing time and effort constraints, our results are consistent with a firm setting where there are no overall effort or time constraints.

2.8 Brith Oucomes

Next we address the question of whether the effect of the incentives for early initiation of prenatal care translated into improved birth outcomes as measured by birth weight, low birth weight, and premature birth. As shown in Figure 2.7 and reported in Table 2.7 we find no effect of the incentives on birth outcomes in either the intervention period or the post-intervention period. There are a number of possible reasons for this. First, the sample could be too small to be able to detect a statistically significant effect on outcomes. However, the point estimates are very small, half of them are negative and they are of similar magnitude to differences between treatment and control groups in the pre-intervention period. Second, given that the results on birth outcomes are obtained from an analysis of a subsample of beneficiaries for whom we were able to merge prenatal care records with hospital medical records, it is possible that the results in Table 2.4 do not hold for this subsample. We therefore replicate the prenatal care analysis using only the subsample of women for whom hospital medical records are available. Overall, we obtain similar results to those obtained

Table 2.6: Cross-Price Effects (Spillover)

	(1)	(2)
	Intervention Period	Post-Intervention Period I
A. Tetanus Vaccine		
Treatment	0.02	-0.02
	(0.08)	(0.05)
Large Sample p-value	0.76	0.62
Wild Bootstrapped p-value	0.75	0.67
Control Group Mean	0.79	0.84
Sample Size	769	1.053
B. Number of visits		
Treatment	0.39	0.51
	(0.33)	(0.58)
Large Sample p-value	0.24	0.38
Wild Bootstrapped p-value	0.27	0.41
Control Group Mean	4.05	4.40
Sample Size	769	1.053

Notes: This table reports LATE estimates of the treatment effect estimated from 2SLS regressions of the dependent variable on actual treatment status instrumented with clinic treatment assignment type. The p-values are for tests of the null that the difference is equal to zero. We present both the p-value computed for large samples and a Wild bootstrapped p-value that is robust in samples with small numbers of clusters (Cameron et al., 2008). Our Wild bootstrap procedure assigns symmetric weights and equal probability after re-sampling residuals (Davidson and Flachaire, 2008) and uses 999 replications. Column (1) reports the results for the sample observed in an 8-month intervention period (May 2010 - December 2010). Column (2) reports the results for the sample observed in the 15-month period following the end of the intervention (January 2011 - March 2012).). Standard errors are in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0$

with the full sample.[29] Third, despite the medical literature and CPG recommendation, it is possible that early initiation of care matters only for a small amount of the general population of pregnant women, such as high-risk patients. High risk patients include, among others, smokers, substance abusers, those with poor medical and pregnancy histories, and those who

29. Results of this analysis are available upon request.

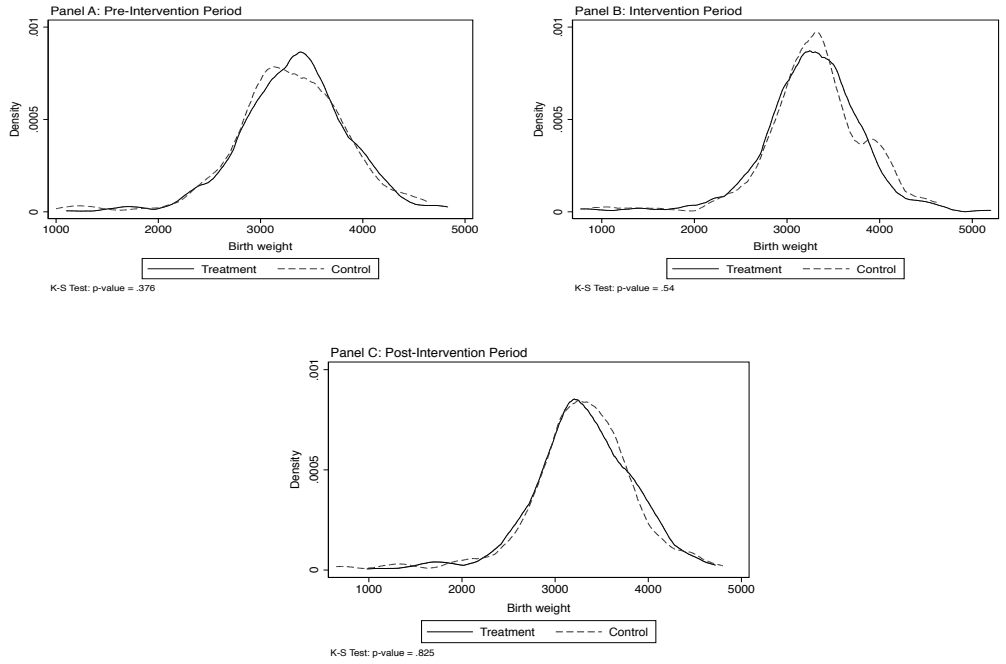

Figure 2.7: Birth Weight Densities

Notes: Densities estimated using an Epanechnikov kernel with optimal bandwidth. P-vales of Kolmogorov-Smirnov tests of equality of distributions between groups reported below figure. Source: Authors' own elaboration based on medical record information system.

start prenatal care very late in their third trimester or only when a problem occurs. It may be that the increase in early initiation of care comes from primarily low-risk mothers who are less likely to benefit from early initiation of care. One would think that it would be easier to persuade low-risk mothers to come a little earlier than to convince high-risk mothers who are reluctant to come for any care at all. In fact, this is consistent with the small reduction in the average weeks pregnant at the time of the first prenatal visit. On average, women in the treatment group initiated prenatal care about 1.5 weeks earlier than women in the control group. Prenatal care may affect birth outcomes by diagnosing and treating illness such as hypertension and gestational diabetes as well as trying to change maternal behavior through promoting activities such as good nutrition, not smoking and not consuming alcohol. If

Table 2.7: Brith Outcomes

	(1)	(2)
	Intervention Period	Post-Intervention Period I
A. Birth Weight		
Treatment	-37.34	25.11
	(48.61)	(40.67)
Large Sample p-value	0.44	0.54
Wild Bootstrapped p-value	0.49	0.51
Control Group Mean	3.304	3.279
Sample Size	555	802
B. Low Birth Weight		
Treatment	0.01	-0.01
	(0.02)	(0.02)
Large Sample p-value	0.63	0.60
Wild Bootstrapped p-value	0.61	0.56
Control Group Mean	0.05	0.06
Sample Size	555	802
C. Premature Birth		
Treatment	0.03	-0.04
	(0.03)	(0.02)
Large Sample p-value	0.31	0.08
Wild Bootstrapped p-value	0.28	0.12
Control Group Mean	0.09	0.12
Sample Size	414	708

Notes: This table reports LATE estimates of the treatment effect estimated from 2SLS regressions of the dependent variable on actual treatment status instrumented with clinic treatment assignment type. The p-values are for tests of the null that the difference is equal to zero. We present both the p-value computed for large samples and a Wild bootstrapped p-value that is robust in samples with small numbers of clusters (Cameron et al., 2008). Our Wild bootstrap procedure assigns symmetric weights and equal probability after re-sampling residuals (Davidson and Flachaire, 2008) and uses 999 replications. Column (1) reports the results for the sample observed in an 8-month intervention period (May 2010 - December 2010). Column (2) reports the results for the sample observed in the 15-month period following the end of the intervention (January 2011 - March 2012).). Standard errors are in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0$

the intervention had induced high-risk women who otherwise would have had their 1st visit much later in the pregnancy, then the incentives may have had a measurable impact on birth outcomes. Hence, while the incentives were effective in increasing early initiation of care, they did not manage to sufficiently affect the group most likely to benefit from it. The solution might be to condition incentives on attending high-risk women, but risk is difficult and expensive to identify and verify and therefore may not be contractible.

2.9 Discussion

In this paper we examined the effects of temporary financial incentives for medical care providers to adopt better quality practices. We used this analysis to investigate whether slow diffusion of better quality practices is driven by perceived low-returns or high fixed costs of adjustment for adopting high return practices. The results suggest that the slow diffusion is driven by high fixed costs as opposed to low returns. We addressed this question in the context of a randomized field experiment in Misiones, Argentina. The intervention randomly allocated a three-fold increase in the fee paid to health facilities for each initial prenatal visit that occurs before week 13 of pregnancy. This premium was implemented for a period of 8 months and then ended. Using data on health services and birth outcomes from medical records, we estimated both the short-term effects of the incentive and whether the effects persisted once the direct monetary compensation disappears. We found that pregnant women who attended clinics in the treatment group were 34% more likely to initiate prenatal care before week 13 and that the higher levels of early initiation of care persisted for at least 24 months after the incentives ended. We also showed that the temporary incentives motivated clinics to design and experiment with new outreach strategies to locate and encourage pregnant women to start care early. For instance, some coordinated with local pharmacies to find out when a woman was late in picking up contraceptive pills, then send community health workers to inquire about last menstruation date, offer instant-read pregnancy tests,

and finally encourage the expectant mothers to start prenatal care quickly. We show that outreach activities for pregnant women doubled in the treatment group. Finally, we provided evidence that in the absence of incentives, it was in the clinics' interest to have provided these outreach services. First, clinic medical directors rank early initiation of care as one of the highest of health priorities among all prenatal care services. Second, outreach activities have been reimbursed at a higher rate than their cost for a long period before the experiment. Likewise, before the temporary fee increase, clinics were paid for each prenatal care service, and 50% of these additional resources were used to pay staff bonuses. As such, the temporary incentives helped to overcome initial costs of experimenting with new outreach strategies for early prenatal care. Once clinics learned what worked best they continue to provide early prenatal care services because they are profitable and valuable. Our results have a number of important policy implications. First, they suggest that temporary incentives may be effective in motivating long-term provider performance at a substantially lower cost than permanent incentives. Second, while we find that incentives are able to motivate changes in clinical practice patterns, we did not find improvements in health outcomes. The monetary incentives that were implemented were not able to sufficiently reach those women for whom early initiation of prenatal care would have the largest health impact. Therefore, incentives may be made more effective by defining ex-ante the population most likely to benefit, and tailoring incentives towards this population. However, tailoring incentives to high risk populations or those most likely to benefit from the services may not be contractible as these characteristics are typically not observable. This is maybe a major limitation of using incentive contracts to improve health outcomes.

CHAPTER 3

CAN SMALL INCENTIVES HAVE LARGE PAYOFFS? HEALTH IMPACTS OF A NATIONAL CONDITIONAL CASH TRANSFER PROGRAM IN BOLIVIA

*In collaboration with Julia Johannsen, Sebastian Martinez, and Cecilia Vidal Fuertes**

3.1 Introduction

Conditional cash transfer (CCT) programs have become widely used policy instruments to promote investments in human capital among the poor in developing countries. Monetary incentives are paid to households conditional upon compliance with requirements, such as visiting health establishments or attending school, intended to improve health, nutrition or education outcomes. Most CCTs have an additional short-term goal of reducing monetary poverty, which is why payments are often equivalent to 10 to 25% of household income and paid in regular quotas (Fiszbein et al., 2009; Stampini and Tornarolli, 2012).

CCT programs have generally shown positive impacts on increasing health care utilization but mixed effects on final health and nutritional outcomes.[1] With relation to maternal and newborn health, past studies have linked CCTs to increases in prenatal visits, skilled attendance at birth, delivery at a health facility, tetanus toxoid vaccination for mothers, and in limited countries to the reduced incidence of low birth weight (Glassman et al., 2013). In child health and nutrition, CCTs have been shown to reduce stunting and prevalence of

∗. Johannsen: Social Protection and Health Division, Inter-American Development Bank, 1300 New York Avenue, NW, Washington DC 20577, jjohannsen [at] iadb.org; Martinez: Office of Strategic Planning and Development Effectiveness, Inter-American Development Bank, 1300 New York Avenue, NW, Washington DC 20577, smartinez [at] iadb.org. Vidal: Unidad de Anlisis de Polticas Sociales y Econmicas (UDAPE), La Paz, Bolivia, cvidal [at] udabpe.gob.bo.

1. See Gertler (2004), Gaarder et al. (2010), Cecchini and Madariaga (2011), and Cecchini and Soares (2015).

underweight in some countries and population subgroups (Fernald et al., 2008; De Brauw et al., 2014; Levy and Ohls, 2010). Furthermore, studies in Mexico (Barham, 2011) and Brazil (Rasella et al., 2013) have shown that CCTs were effective in reducing infant and child mortality. Yet most studies of anti-poverty CCTs do not separate health effects derived from increased utilization of health services from other components of the CCT, including sizable income transfers and parenting courses which can affect health through increased consumption of food, medications, and health and nutritional practices in the household.

An alternative CCT model focuses on improving health outcomes by stimulating the demand for specific health services through small directed transfers. The Janani Suraksha Yojana program in India, for example, pays mothers between $13 to $31 USD conditional on delivery in a health center, and has been shown to increase institutional deliveries and lower perinatal and neonatal deaths (Lim et al., 2010; Randive et al., 2013). With the exception of the program studied here in Bolivia, this model of CCT has not been widely applied in the Latin America and Caribbean region.

We study the effects of a national conditional cash transfer program in Bolivia, the Bono Juana Azurduy (BJA), that pays pregnant women and mothers of children under two years old between $7 and $18 per visit for prenatal checkups, skilled birth attendance and preventive health care checkups for children. Using a variety of identification strategies, we analyze the effects of the program on health services utilization of prenatal, intrapartum and postpartum care, routine checkups for children, and associated health outcomes including stillbirth and child survival, low birth weight, height, weight, and anemia. We find that the program effectively stimulates the demand for health services, has mixed results on final health outcomes, and is overall highly cost-effective to a large extent because of the modest amounts of the transfers relative to other CCT programs operating in the region.

Despite the provision of free or low-cost maternal and child healthcare in many countries, the utilization of preventive services often remains below recommended levels, particularly

for poor, rural and indigenous populations (Mills, 2014)).[2] In countries like Bolivia, where four out of ten pregnant women have their first contact with the health system in the second trimester of pregnancy or later,[3] failures to detect high-risk pregnancies or to respond to obstetric and neonatal emergencies can have deleterious effects on the health and survival of mothers and their newborns. Similarly, inadequate preventive care for young children, including immunizations and nutrition supplementation, can negatively affect a child's long-term health and development (Strauss and Dietz, 1998; Tamura et al., 2002; Martorell et al., 2010). Consistent with other countries, many Bolivian mothers shy away from free health services for a host of reasons, including perceptions on quality of care and distrust, access barriers such as distance to health facilities and transportation costs, opportunity costs at home and work, gender barriers, and time inconsistencies that lead patients to delay or postpone health seeking behavior (Thaddeus and Maine, 1994; Ensor and Cooper, 2004; Thaler and Sunstein, 2009; Banerjee and Duflo, 2011). To the extent that the under-utilization of health services generates private and social costs from preventable illness and mortality, identifying cost-effective mechanisms to stimulate the demand for health care is a pressing public policy question.

3.2 Context and Intervention

3.2.1 Country context

During the past decades, Bolivia has experienced significant improvements in its population health and nutrition indicators; however, compared to other Latin American countries, health

2. See also the report "Universal Health Coverage report" by the World Health Organization in *http* : *//www.who.int/universal_health_coverage/en/*. Access 05/11/2015.

3. Authors' calculation using ESNUT 2012. In Bolivia, the first prenatal care visit, on average, takes place at 13.6 weeks of pregnancy.

indicators in Bolivia continue to be among the worst in the region (Organization, 2013).[4] According to Demographic Health Surveys (DHS) available for the country, approximately four of every ten under-five deaths occurred during the first month of life in 2008. Moreover, neonatal mortality reported very little progress between 2003 and 2008, remaining at 27 deaths per 1,000 live births. Maternal mortality rate in 2003 was as high as 229 deaths per 100,000 live births according to the DHS of that year.[5] In terms of child nutrition, despite large improvements in the past two decades, DHS data from 2008 show that 26% of children under 3 years old suffered from chronic malnutrition; and that stunting rates of children in rural almost doubles the rate in urban areas. Low coverage of basic maternal and child health services may explain these numbers. Prior to 2009, 71% of total births in Bolivia were delivered by skilled health personnel and 72% of pregnant women had at least four antenatal care visits during their pregnancy (see Table 3.1). In rural areas, these numbers were significantly lower, reaching 51% and 60%, respectively. Table 3.1 shows that these indicators are much higher relative to the Latin American region, where 94% of deliveries are attended by skilled personnel and 86% of pregnant women had at least four prenatal medical visits as recommended by Organization (2006).

3.2.2 The Bono Juana Azurduy Program

Since 1997, uninsured pregnant women and children under five years old in Bolivia are covered by the Universal Maternal and Child Insurance (SUMI by its Spanish acronym). SUMI provides a free basic health care package to improve access and coverage of care. To incentivize demand for maternal and child-care health services provided by SUMI, in May of 2009 the government launched a nation-wide conditional cash transfer program, the Bono

4. Mortality of children under 5 years old dropped from 115.6 per 1,000 live births in 1994 to 63 per 1,000 live births in 2008, whereas the WHO estimated average for the Latin America and the Caribbean region in 2011 was 16 per 1,000 live births.

5. The official maternal mortality figure in Bolivia has not been updated since 2003.

Table 3.1: Coverage indicators for utilization of maternal and child health services before the BJA

Indicator	Bolivia (Urban/Rural) 2008* (%)	Average LAC** (%) (2011)
Use of birth control methods	61 (66/53)	74
Percentage with at least 4 pre-natal medical visits	72 (82/60)	86
Births delivered by qualified personnel	71 (88/51)	94
Immunization rate for BCG Vaccine (18-29 months)	98 (99/98)	–
Immunization rate for DPT vaccine (18-29 months)	86 (85/87)	92

Notes: Sources: (*) Demographic and Health Survey 2008. (**) WHO (2013).

Juana Azurduy (BJA). BJA incentivizes the use of maternal and child health services by pregnant women and children under two years old through the payment of cash transfers that are conditioned on the use of select clinically recommended preventive services. Enrollment in the program is voluntary and all pregnant women and under-two-year-olds not covered by the social security system are eligible to enroll. According to Census estimates, in 2012 over 82% of women and children in the country were eligible to enroll.

The program's conditionalities, or "co-responsibilities", and associated payments are detailed in Table 3.2. BJA pays pregnant women $50 Bs ($7 USD) for each prenatal visit up to a maximum of four visits, and $125 Bs ($17 USD) for skilled birth attendance, whether delivery takes place at a health facility or at home assisted by qualified health personnel. For children under two years old, the program pays $120 Bs ($18 USD) for each bi-monthly integral health check-up in the first 2 years of life, up to a maximum of 12 visits. With full compliance of these conditionalities, covering 9 months of pregnancy and the initial 24 months of the child's life, the maximum cumulative transfer amounts to $1,820 Bs ($261 USD) over a 33 month period. This total amount represented approximately 1.3 minimum wages in 2014.

BJA's conditionality structure differs from most anti-poverty CCT programs in the LAC

Table 3.2: Co-responsibilities and amounts in the BJA

Corresponsability	Number	Amount ($Bs/$USD)	Maximum ($Bs/$USD)
Women during pregnancy:			
Prenatal medical visit	4	50 / 7	200 / 28
Delivery assisted by qualified personnel	1	120 / 17	120 / 17
Total benefits from pregnancy			320 / 45
Children under 2 years old:			
Preventive health care visit	12	125 / 18	1.500 / 216
Total benefits from children			1.500 / 216
Total benefits over entire cycle (33 months)			1.820 / 261

Notes: Authors' own elaboration.

region in a number of ways. First, the program excludes education and general poverty reduction components in its payment structure and conditionality design. Instead, the program focuses exclusively on human capital accumulation in terms of maternal and child health. Furthermore, the program includes a clearly defined exit strategy based on the age limit of two years and beyond which children are not eligible. BJA is also universally available to all uninsured women and children and does not target specific groups of poor or vulnerable households.[6] In addition, the program pays transfers individually for each eligible health visit completed by mother or child according to the specific amount related to the conditionality that is due, in contrast to regular monthly or bi-monthly payments in most CCT programs. Finally, the total transfer amount is small in terms of household consumption. We estimate that cumulative transfers are equivalent to less than 1% of average per-capita consumption over the 33-month period, compared to other programs in the region that can range from 7% to 31% of consumption (Fiszbein et al., 2009).[7] As such, we posit that any observed health effects of BJA are more likely to be the result of increased health care rather

6. In fact, BJA beneficiaries are nearly evenly distributed between income quintiles.

7. Fiszbein et al. (2009) show that the average transfer in CCTs implemented in LACs is approximately 17% of average household consumption.

than an income effect.

3.2.3 Enrollment in the BJA

BJA started enrolling women and children on May 11th of 2009. The first payment was delivered on May 27th of the same year. Operational rules establish that enrollment should be done at the public health center that is closest to the beneficiary's home. Although the BJA benefits children until they turn two years old, they are required to be younger than 12 months at the time of enrollment, in order to guarantee a minimum exposure of 12 months to the program. As a pre-condition for enrollment, an identity card and birth certificate must be presented for the pregnant women and children, respectively, in addition to a pregnancy test for pregnant women. Figure 3.1 and Figure 3.2 show the evolution of enrollment rates in the program for pregnant women and under-one-year-olds obtained from the Health and Nutrition Evaluation Survey 2012 (ESNUT 2012), and BJA administrative records.[8] On average, the enrollment rate of eligible women was approximately 33% between 2009 and 2012, with a decreasing trend over this period. The enrollment rate of children during the same period was approximately 52%. According to ESNUT 2012, amongst non-enrolled eligible mothers, the main reasons for not enrolling are the lack of information about the program (27.5%), not having the required legal documents at the moment of enrollment (19.9%) and time costs associated to long queues or long trips to health facilities (20.3%).

8. Enrollment rates are based on retrospective data from the Health and Nutrition Evaluation Survey (ESNUT) 2012 survey and BJA enrollment records (as the numerator), and official population projections (pregnancies and one-year olds) for the denominator.

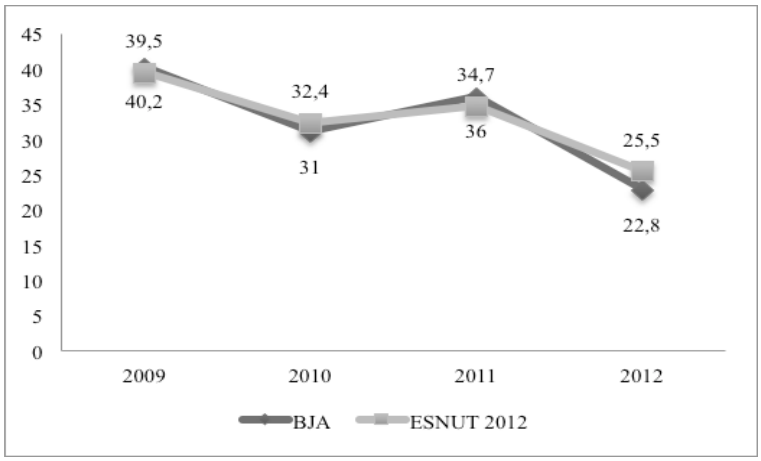

Figure 3.1: Enrollment rate in the BJA of pregnancies by year of pregnancy

Notes: Source is ESNUT 2012 survey and BJA enrollment records. For the enrollment rate using BJA records, we use official population projections as the denominator of eligible population.

3.2.4 Mechanisms to improve health outcomes embedded in the BJA Program

One of the main outcomes that we study in this paper is the effect of the BJA on the rate of mortality at birth.[9] One of the primary mechanisms through which the BJA may affect stillbirths is by improving prenatal care utilization, i.e. increasing the number of average visits during pregnancy and increasing early initiation of prenatal care. Early initiation of prenatal care is part of standard training in nursing schools throughout the world (Organization, 2006) and has been linked to positive maternal and newborn health outcomes (Carroli et al., 2001; Campbell and Graham, 2006b). In particular, early detection of medical conditions such as maternal infections or anemia in the period in which the fetus is most at risk can improve outcomes at birth, such as low birth weight, prematurity and mortality at birth (Carroli et al., 2001; Hawkes et al., 2013). Early prenatal care also allows providers

9. Mortality at birth can include stillbirths, abortions, or any other definition of death during the first 24 hours after birth.

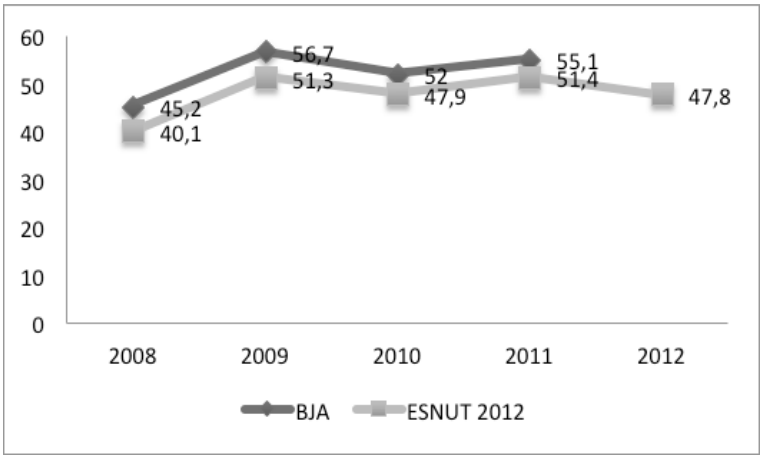

Figure 3.2: Enrollment rate in the BJA of children by year of birth

Notes: Source is ESNUT 2012 survey and BJA enrollment records. For the enrollment rate using BJA records, we use official population projections as the denominator of eligible population.

to advise mothers on proper prenatal nutrition and prevention activities (Phavichitr and Catto-Smith, 2003). The number of medical visits while pregnant may also be associated to birth outcomes since these provide an opportunity to monitor complications that may appear and reinforce healthy behavior throughout the complete period of pregnancy. In addition to preventive prenatal care, the presence of skilled health professionals at delivery can have positive outcomes on mortality within the very few days since birth (Moss et al., 2002). We test for this by studying how the BJA affected the number of hospital deliveries relative to home-based deliveries by skilled professionals.

Another group of health outcomes that we study in this paper are related to child nutrition. Medical visits incentivized by the BJA could improve nutrition through at least two channels. First, the BJA not only incentivizes the number of medical visits but also that these visits are completed at regular intervals during the first two years. Investments at earlier stages of growth can have larger returns (Heckman and Mosso, 2014). In particular, medical visits can educate mothers about how to best care for their children while also

help to detect any health complications that affect a child's growth (Thapar and Sanderson, 2004). Second, in Bolivia children are entitled to a free nutritional supplement administered at health facilities during medical visits. The supplement contains iron, zinc, vitamins A and C, and folic acid, which can have a positive effect on reducing the prevalence of anemia (Lopez et al., 2015).

3.3 Data

3.3.1 National Health Services Information (SNIS) and Census data

Our primary outcome of neonatal deaths comes from the National Health Services Information (SNIS by its Spanish acronym), a national registry of information to which municipalities must report on different indicators of health services utilization and health status of their population. The information is interactively available on the Ministry of Health website.[10] We downloaded information for each municipality on the number of total births, live and stillbirths. Municipalities consolidate information on the result of each birth on a monthly basis. To enter this information, a midwife, medic or a qualified health professional must have attended the birth at a hospital or at home. There is no clear information on what is considered a death, so this classification may include abortions. Neonatal death (referring to the first 7 days of life) does not form part of this registry. Therefore, we define the variable to be a count of stillbirths due to all causes, without specifying the exact reason of the pregnancy termination.

Our second source of information is the Population Census of 2012, which we use to validate our analysis using SNIS data. The principal goal of the Census is to update information on the number of people and households, and their distribution over the country. It also provides information on demographic characteristics, economic well-being and housing of

10. *http* : *//www.minsalud.gob.bo/*. Access 05/30/2015.

each person or households covered. It is implemented every 10 years, and in 2012 it covered a total of 10,027,254 inhabitants.

To evaluate the effects of the BJA on mortality at birth using SNIS data we construct a municipal level panel of stillbirths per 1,000 live births from 2005 to 2012. The resulting dataset includes 327 municipalities and 2,616 municipality-year observations. The data is reliable only for births attended by a health professional in hospitals or at home. For some municipality-year observations the system reports zero stillbirths or live births. This may reflect intermittent reporting from some municipalities in particular years and not necessarily a true absence of births or stillbirths. We argue in section 3.4 that our identification strategy for the effect of the BJA on the rate of stillbirths is robust to the differential rates of birth reporting across municipalities in the country.

To complement the SNIS analysis described above, we use aggregated 2012 census data on age-cohort population sizes to study the effect of the program on child survival. For each municipality we count the number of children ages 0 to 6 separately and treat this data as a panel of municipalities, where each observation represents a municipality-year birth cohort of surviving children.[11] Children ages 5 or 6 years in 2012 were never exposed to the BJA, since they were born two years before its implementation. Hence, the BJA should have no effect on the cross-municipality differences in cohort-size for these age-cohorts, but could have had an effect on younger cohorts by increasing child survival. We develop this idea further in section 3.4.

Finally, the SNIS data indicates whether each birth recorded in the sample was delivered at the domicile of the mother or at the hospital. This is an important outcome since it could explain whether the BJA incentivized women to deliver at hospitals in an environment better equipped to attend last-minute complications during deliveries. We construct the ratio of total deliveries at the domicile to total deliveries at the hospital in each municipality and

11. A similar approach was implemented by Jayachandran (2009) to study the effect of wildfires on early life mortality in Indonesia.

year.

3.3.2 Health and Nutrition Evaluation Survey (ESNUT 2012)

The Health and Nutrition Evaluation Survey (ESNUT, by its Spanish Acronym) 2012 is a nationally representative household survey implemented by the Plurinational State of Bolivia to provide information for the evaluation of national health and nutrition programs, including the Zero Malnutrition Program and the BJA. The survey provides information about the health and nutritional status of the Bolivian population, as well as allowing the construction of health-system coverage indicators, with a strong emphasis on maternal and child health care. The sample design allows disaggregation by urban and rural areas, as well as by ecological regions (highlands, valleys and lowlands). It considers a multistage probabilistic sample selection using the 2001 Census as the sampling frame for the selection of primary sampling units (PSU). The survey provides sampling weights to adjust for different selection probabilities. The full sample covers 8,433 households (2,456 urban and 5,977 rural) in 424 PSUs, and is representative at regional (highlands, valleys and lowlands) and urban/rural levels.

The ESNUT 2012 includes a basic demographic household questionnaire, a questionnaire for women in reproductive age (14 to 49 years) and a questionnaire for children under 5 years old living in the household. For all women interviewed that had at least one pregnancy since January 2007, the survey collected retrospective information on every pregnancy, including information on the number of antenatal care visits, birth outcomes and post-partum visits. For children under 5 years old, the survey collected information for all medical visits, immunization, current nutritional status and anthropometric measures of height and weight, as well as hemoglobin levels for children 3 months or older.

In addition, the ESNUT 2012 asked retrospective questions about participation in the BJA program for all pregnancies and children, allowing us to identify beneficiary households

in the survey. The sample of ESNUT 2012 households used for the analysis excludes mothers covered by social security, about 16% of the total, and includes mothers with at least two children and/or pregnancies, resulting in an analysis sub-sample of 5,518 pregnancies and 5,517 children nation-wide.

3.4 Empirical Strategy

As discussed above, BJA was implemented nation-wide in 2009, targeting pregnant women and children younger than 12 months at the time of enrollment. Yet enrollment rates varied substantially across the country, and only about one in three eligible pregnancies and one in two eligible children were enrolled in the program during the study period. Our quasi-experimental identification strategies identify the impacts of BJA through arguably exogenous variation in program eligibility and enrollment over time and across space. Our empirical strategies are implemented depending on the outcome of interest and the data source. Below, we divide them accordingly.

3.4.1 Municipality Level Outcomes: Rate of Stillbirths and Hospital Deliveries

We estimate the effect of BJA coverage within a municipality on the rate of stillbirths using the SNIS data. We construct the number of still-births (numerator) per 1,000 live births (denominator) in each municipality for each year between 2005 and 2012. The treatment variable is the percentage of eligible women enrolled in the BJA in each municipality for each year between 2009 and 2012.[12] We estimate the following regression:

$$Y_{j,t} = \phi_t + \phi_j + \delta_1 Enroll_{j,t} + \delta_2 Enroll_{j,t-1} + X'_{j,t}\gamma + \varepsilon_{j,t} \tag{3.1}$$

12. For years 2005 to 2008 the enrollment rate is equal to zero.

Where, $Y_{j,t}$ is the rate of stillbirths for municipality j at year t; $Enroll_{j,t}$ is the ratio of number of enrolled women (numerator) to total number of eligible women (denominator) in municipality j at year t; $X_{j,t}$ is a vector of controls; and ϕ_j, ϕ_t, and $\varepsilon_{j,t}$ are municipality fixed effects, time fixed effects, and unobservable characteristics that vary with municipality and time, respectively.

We include a one period lag in the enrollment rate, $Enroll_{j,t-1}$, since women that gave birth in the first months of a given year, were exposed to the BJA during the previous year for most of their pregnancy.[13] We include a binary indicator for each year to control for shocks that are common to all municipalities, such as changes in medical guidelines. Municipality fixed effects control for unobserved variables specific to the municipality that are fixed over time, such as altitude and weather, among others. As such, the source of variation for the identification of δ_1 and δ_2 is the variation in the enrollment rate over time within the same municipality.

The key identifying assumption in (3.1) is that the change in the outcome observed in municipalities with high enrollment rates would have been the same, had they experimented lower levels of enrollment, as those that actually experimented lower levels of enrollment. And vice-versa. Although this assumption is not testable, commonly known as "parallel-trends" in potential outcomes, we are able to explore whether the pre-intervention trends in outcomes were similar across municipalities with different take-up levels. If the trends are the same in the pre-intervention period, they are more likely to have been the same in the intervention period. Figure 3.3 shows the evolution of the rate of stillbirths per 1,000 live births in our time period. The y-axis shows the rate of stillbirths in deviations from the group-average, and the x-axis shows calendar years. We divide our sample into municipalities whose enrollment rate in year 2009, the first year of the BJA, was above or below the median enrollment rate of the country and plot the rate of stillbirths over time for these two groups.

13. Another possible effect of including the lag is that it also captures the effect of exposure to the program during first months of in-utero development for children born in the same year.

Figure 3.3 provides a graphical representation of what we aim to estimate in (3.1). The plot shows that, with the exception of year 2005, the trends in the rate of stillbirths are similar between the groups in the years previous to the program, a necessary condition for the identification of the treatment parameters in (3.1). In addition, the stillbirth rate in both groups, hence nationally, has decreased over time, however once the BJA starts the stillbirth rate decreases at a much faster rate in municipalities with higher enrollment rates in the program.

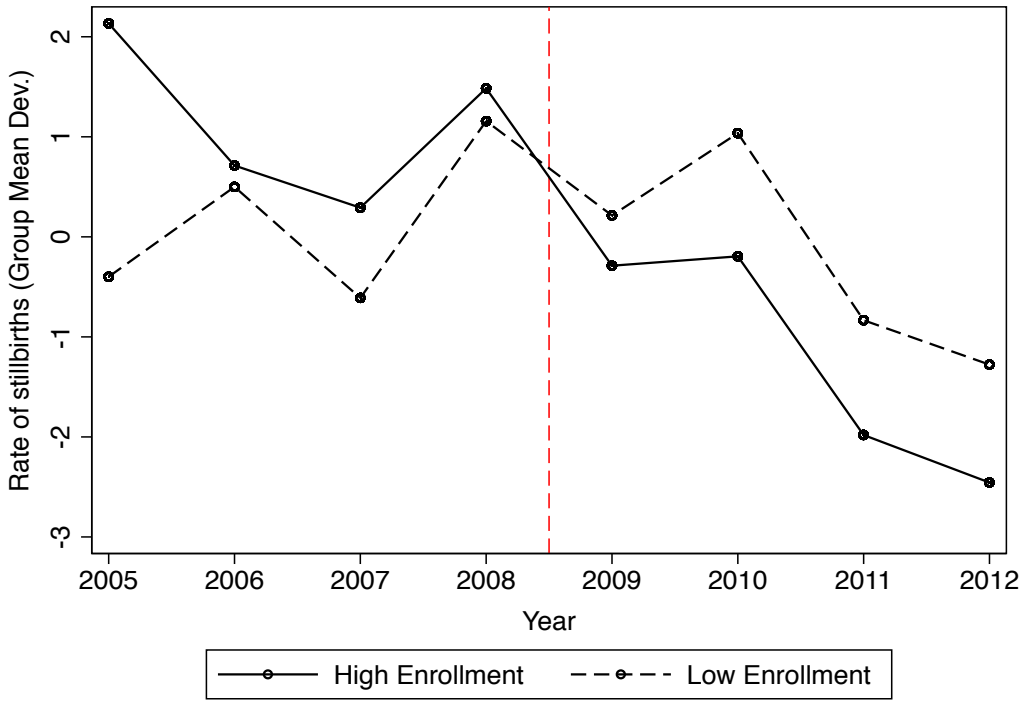

Figure 3.3: Trends in the rate of stillbirths for municipalities with enrollment rates above the median (High) and below the median (Low) enrollment rate in year 2009

Notes: The Figure shows the evolution of the rate of stillbirths per 1,000 live births in our time period. The y-axis measures the rate in deviations from the average and the x-axis shows calendar years. We divide our sample into municipalities whose enrollment rate in year 2009 was above or below the median enrollment rate of the country and plot the rate of stillbirths over time for these two groups. In this graph we drop 2005 and observations with 0 rate of stillbirths to smooth trends.

To test the "parallel trends" assumption across municipalities for the years previous to

the program, we run the following regression using data for years 2005 to 2008:

$$Y_{j,t} = \alpha + \sum_{t=2006}^{2008} \beta_t Year_t + \sum_{t=2006}^{2008} \gamma_t Year_t Enroll_{j,2009} + \phi_j + \mu_{j,t} \tag{3.2}$$

In this regression, the γ_t coefficients capture the correlation between the enrollment rate in year 2009, at the start of the program, and any change in the outcome in years 2006 through 2008 with respect to year 2005. We include time and municipality fixed effects. A test for the joint significance of the parameters γ_t provides a test for the parallel trends condition necessary for the identification of δ_1 and δ_2 in (3.1). The results are shown in the first column of Table 3.4. The change in the rate of stillbirths before year 2009 is uncorrelated with the enrollment rate at the beginning of the BJA. The joint test, provided at the bottom of the Table, shows a p-value of 0.353.

3.4.2 Pregnancy Outcomes: Prenatal Care Utilization

The ESNUT 2012 collects information on utilization of health services for every pregnancy experienced by women living in the household with a pregnancy since 2007. We select BJA-eligible women with at least two children and for which at least one child was born before the start of the program and at least one child was born after the start of the program. We select two observations for each mother, one that corresponds to the pregnancy of the last child born before June 1st of 2009, when the BJA started, and one that corresponds to the pregnancy of the first child born after this date. This allows comparing, for the same mother, one pregnancy that ended before the BJA, and one pregnancy that ended after the BJA. The pregnancy for the child born after June 1st, 2009, was eligible to enroll in the BJA.

Consider the following regression:

$$Y_{is} = \alpha + \delta D_{is} + X'_{is}\beta + \eta_i + \gamma_s + \varepsilon_{is} \tag{3.3}$$

Where subscript s refers to a pregnancy and i refers to a mother. Y_{is} is an outcome, such as weeks pregnant at the first prenatal visit, for pregnancy s of women i; D_{is} is a binary indicator for whether pregnancy s of mother i was enrolled in the BJA; X_{is} is a vector of controls that are observed and can be fixed or vary across different pregnancies for the same mother (e.g. education or order of birth); η_i are unobserved fixed variables for mothers (e.g. parental skills); γ_s are unobserved fixed variables for pregnancies (e.g. unobserved risk measures); and ε_{is} are unobservable variables that are allowed to vary within a mother across her different pregnancies. Cross-sectional estimation of equation (3.3) by OLS will generally give biased estimates of the treatment parameter, δ, because unobserved components of η_i and γ_s may influence the probability of treatment take-up and pregnancy outcomes.

To eliminate this potential source of bias we estimate a fixed-effects model comparing the same outcome between treated and non-treated pregnancies of the same mother, thus eliminating the fixed unobserved component, η_i , in equation (3.3). Program impacts are estimated in the sub-sample of mothers that have two or more pregnancies. The variation to identify D_{is} is induced by mothers with at least one pregnancy or child ineligible and another one eligible for the BJA, based on the program's age eligibility restriction, which defines that only children under one year of age could enroll in May 2009. This allows for the selection into treatment to be mainly due to the eligibility rule, which is independent of

the potential outcomes of each observation. We estimate the following linear regression:[14]

$$\tilde{Y}_{is} = \delta \tilde{D}_{is} + \tilde{X}'_{is}\beta + \phi_s + \tilde{\varepsilon}_{is} \tag{3.4}$$

Where the transformation of a variable Z_{is} to $\tilde{Z}_{is}$ indicates that the variables deviate from the group mean. Assuming that the treatment status is uncorrelated with the new unobservable component, $\tilde{\varepsilon}_{is}$, we can estimate (3.4) with an OLS regression and obtain an unbiased estimate of the treatment effect. The assumption would be violated if mothers behave systematically different from one pregnancy or child to another. However, in the context of BJA, we argue that the variation in the treatment status for the same mother is due mostly to exogenous program eligibility rules (namely the implementation of the program for children under 12 months of age in May 2009) rather than to unobserved components that change from one pregnancy to another. However, conditional on a pregnancy or child being eligible, mothers still make the choice of enrolling to the BJA during their pregnancy or enrolling their children in the program. To address this issue we include controls for order of birth, age of the mother at delivery, sex of the child and cohort of birth, to control for common time trends and factors such as gains in parenting experience over subsequent pregnancies or children.[15] Furthermore, the average time between pregnancies or children in our sample is approximately 2.5 years, which constrains the likely variation in household

14. We estimate OLS regressions for both continuous and discrete outcomes. For the latter, index models in fixed effects settings may be computationally intractable and impose assumptions on the functional forms of the data that could lead to worse bias than Linear Probability Models. See a discussion of this by Steve Pischke in the online blog of his book "Mostly Harmless Econometrics". $http://www.mostlyharmlesseconometrics.com/2012/07/probit-better-than-lpm/$. A main restriction on running linear models is that the predicted probabilities may be outside the unit interval. In most cases, estimating equation (3.4) without vectors $\tilde{X}_{is}$ provides estimates of the treatment effects close to those obtained when such vector is included. As such, the "uncontrolled" version of (3.4) is a fully saturated model in which the predicted probabilities are constrained to the unit interval by construction. In addition, our main purpose is to estimate the partial effect on the response probability, averaged across the distribution in the sample, as opposed to the effect on particular values of the distribution. In this case, the linear probability model is appropriate. See page 454 - 457 in Wooldridge (2010). The heteroskedasticity problem for variance estimation is addressed by using heteroskedasticity-consistent robust standard error estimates.

15. For a similar analysis see Salm and Schunk (2012).

environments, especially if such changes must differ systematically between treated and untreated observations to pose a risk to the identification strategy. Furthermore, albeit at a different scale and for a different sample, the analysis at the municipality level in the previous section supports the parallel trend assumption also needed in this analysis.[16]

3.4.3 Children Outcomes: Utilization of Health Services and Health Outcomes

The program started in the middle of May of 2009 and it was implemented nationwide. The eligibility criteria for children only included children of 12 months of age or below at the moment of enrollment. Any child older than 1 year old was not eligible. This provides a natural discontinuity in the eligibility rule based on a child's age at the moment of enrollment. To estimate the effects of the BJA on the utilization of health services by children we employ a regression discontinuity (RD) design (see Lee and Lemieux, 2010).

To explore how likely are households complying with the age-specific eligibility rule, we use the ESNUT 2012 and study the relation between date of birth of each child in our sample and enrollment rates in the BJA. We use June 1st of year 2008, as the cutoff date for birth date.[17] A graphical representation of this relation is shown in Panel A - Figure 3.4. The x-axis is the date of birth of children and the y-axis shows the dependent variables mentioned above. The horizontal line in the middle of the graph is fixed at June 1st of year 2008, exactly one year before the BJA started. Children that are born before the cut-off date are grouped to the left of the vertical line, and those born after are grouped to the right. Each bin corresponds to a week and we use 72 days before and after the cutoff date as the bandwidth. Children born before June 1st 2008 were too old to enroll. The graph shows the

16. We also attempted to do the parallel trends test using mothers with more than two children in pre treatment periods. However the sample size are small.

17. In particular, we use each child's date of birth to construct the difference in weeks to June 1st of year 2008. Within each week we plot the mean enrollment rate.

take-up rate is very low or zero as we move further to the left in the x-axis. To the right of the vertical line, the take-up rate increases substantially. Our RD estimates show that the take-up rate increases by more than 30 percentage points at the cut-off of eligibility. The Figure also shows that the change in BJA take-up is not sharp at the cut-off, i.e. take-up is not 100% after the cutoff date. To account for this, we implement a "fuzzy" RD design.

Formally, let B_i be date of birth of each child i and consider a cut-off c, set at June 1st of year 2008. Let D_i be an indicator variable for whether child i is enrolled in the BJA, and Z_i, be an indicator variable equal to one when the child's birth date is equal or exceeds the cut-off of June 1st in year 2008, indexed by c, i.e. $Z_i = 1(B_i \geq c)$. The two equations we estimate in the "fuzzy" RD are:

$$Y_i = \beta_0 + \beta_1 D_i + f(B_i) + \varepsilon_i \tag{3.5}$$

$$D_i = \gamma_0 + \gamma_1 Z_i + f(B_i) + \mu_i \tag{3.6}$$

Where Y_i is an outcome, such as number of medical visits that occur during months 12 to 24 of age, for each child. The function $f(B_i)$ is a smooth function of the birth date, which is the forcing variable that determines whether a child is eligible or note to the BJA. It captures smooth, seasonal effects of birth dates on outcomes. The central assumption underlying the RD design is that we have correctly specified $f(B_i)$. We estimate a local (kernel-weighted) linear regression to the left and right limits of the discontinuity choosing different bandwidths. We work with the sample of children that are two years old or more at the moment of the survey. This restriction drops all children that may still have on-going treatment since they can be covered by the BJA until they are 2 years old.

In practice, a "fuzzy" RD exploits discontinuities in the probability of treatment conditional on a covariate. The discontinuity is used as an instrumental variable for actual take-up of the program and the usual assumptions of Instrumental Variable (IV) estimation are needed (see Hahn et al., 2001). Hence, the parameter that we estimate using the "fuzzy"

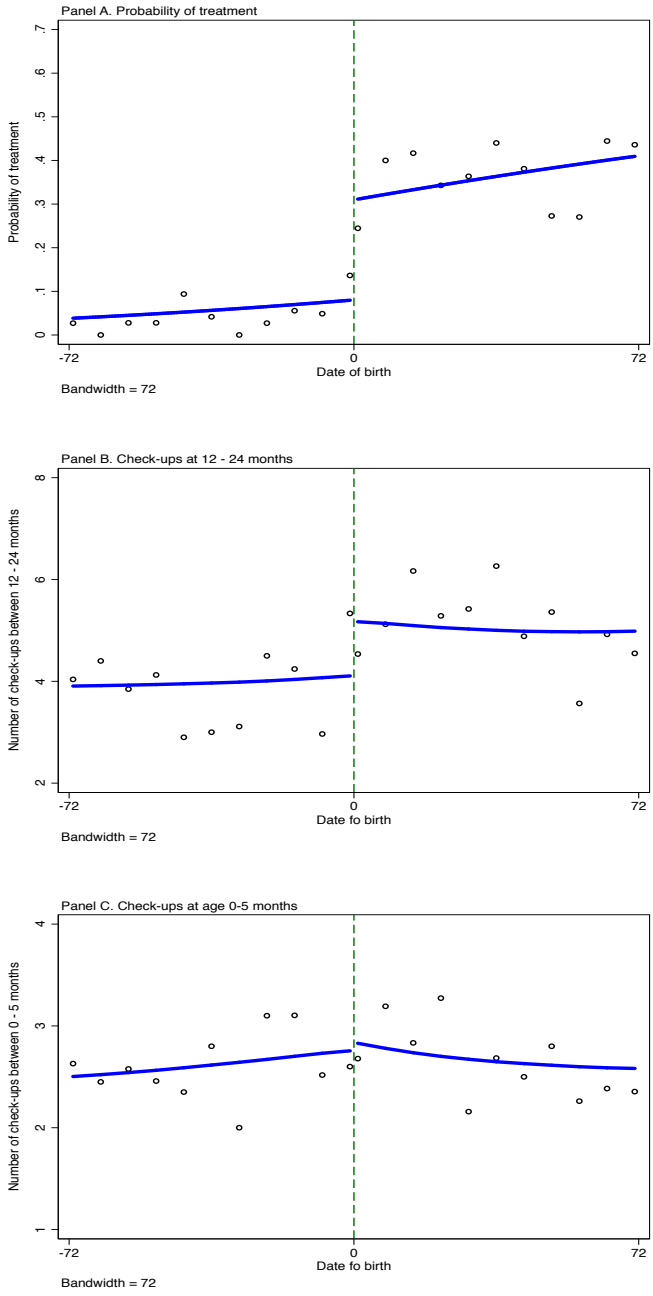

Figure 3.4: RD-Graphical Analysis. Effects of the eligibility rule on treatment take-up, number of check-ups at 12 - 24 months of age, and placebo test for number of check-ups at 0-5 months of age.

Notes: The Figure is a graphical representation of the effect of the age specific eligibility rule on treatment take up (Panel A), the number of medical visits between months 12 to 23 of age (Panel B), and the number of visits between months 0 to 5 of age (Panel C). The x-axis is the date of birth of children and the y-axis shows the dependent variables mentioned above. The horizontal line in the middle of the graph is fixed at May of year 2010, the month when the BJA started.

RD is the Local Average Treatment Effect (LATE), which is the effect on those observations that comply with the instrument, i.e. children that are moved into the treatment because of the age-specific rule (see Guido W. Imbens, 1994).

Finally, the identification strategy in RD relies on the assumption that observations are randomly assigned to the treatment at the cut-off. An indirect test for this assumption is to check the balance across different covariates within the cut-off. We test for this by running a local linear regression version of (3.5) for different demographics and covariates that should not vary with treatment status. All results are commented in section 3.5.

In addition to the RD results we use the longitudinal structure of the ESNUT 2012 data and run the same analysis as for pregnancy outcomes. In this case we select two children for each mother, one eligible and the other non-eligible, and estimate the effect of the BJA using the mother fixed effects approach. The same assumptions apply. The effects should be different than the RD analysis in that we are using different samples and estimating different treatment parameters. Finally, the validity of our results depends, at least in part, on the fact that children enrolled in the BJA are not different, other than in their treatment status, from non-enrolled children. Table 3.2 and 3.3 in the Appendix show descriptive statistics for the subsamples of children used in the fixed effect analyses. Columns 1 and 2 in Table 3.4 shows that households whose children are enrolled in the program include younger, richer and more educated mothers. Even though all these characteristics are fixed and hence controlled for in the fixed effect analyses, they are informative about the differences in composition across households that should be considered in the interpretation of our results. Our results are not externally valid and they should be interpreted as the effect within households similar to the sample, i.e. households with mothers with more than two children during the time period of the survey.

3.5 Impacts of the BJA

3.5.1 Municipality Level Outcomes: Rate of Stillbirths

Regression (3.1) is estimated controlling for a binary indicator for each period in our sample to control for common time trends, ϕ_t, and including the number of health facilities and payment centrals of the BJA per 1,000 habitants as controls in X_{jt}. We also drop municipality-year observations that have a stillbirth rate higher than 300 per 1,000 (2 observations), and exclude municipalities with more than 250,000 habitants, which are very different from the rest of the country. The final sample size consists of 321 municipalities of the 327 available in the country, for a total municipality-year sample size of 2,566. We run the regressions weighting each observation by the population size of the municipality in the Bolivian Census 2012 to adjust for the heteroskedasticity induced to the standard errors when using aggregated data.[18]

Table 3.3 shows the results of different specifications of equation (3.1). The first column shows that municipalities with higher enrollment rates also experienced a steeper decline in the rate of stillbirths. The estimated coefficient shows that a 1-percentage point increase in the lagged rate of enrollment in the BJA is associated with a reduction of 0.06 stillbirths per 1,000 live births. While our results only show that the lagged value of enrollment rate at the municipality level is statistically significant, a joint test for both coefficients shows that lagged and contemporaneous enrollment rates are jointly significant at conventional levels (p-value=0.03). Adding both coefficients, contemporaneous and lagged enrollment rate, a 1 ppt. increase in the enrollment rate of the BJA is associated with a reduction of 0.083 stillbirths per 1,000 live births. To approximate this association in an average municipality under current enrollment rates, we convert these results for the average enrollment rate of the sample. The average over time across municipalities is 32%. As such, an average

18. See Solon et al. (2013).

municipality experienced a decline of 0.32×8.3=2.7 stillbirths per 1,000 live births, which corresponds to a 12.4% decline with respect to the average rate of stillbirths at baseline. To assess the heterogeneity of these findings, we estimate the same regression for the sub-sample of municipalities with poverty rates above and below the median poverty rate of the sample. The results show that the absolute decline in stillbirths is higher in poorer municipalities, where it amounts to a relative reduction of 21% in the rate of stillbirths with respect to the baseline rate for this sub-sample. We find that, although the coefficients are also negative, there are no significant effects on the sub-sample of municipalities below the median poverty rate.

Table 3.3: Effect of BJA Intensity on the Rate of Stillbirths at the Municipality Level

Dep. Var.: Rate of Stillbirths	Main Results		Below Med. Poverty	Above Med. Poverty
	(1)	(2)	(3)	(4)
Enrollment rate t	-2.167	-2.093	-6.936	0.185
	(2.361)	(2.385)	(4.506)	(2.728)
Enrollment rate t-1	-6.096**	-5.989**	-9.037*	-4.814
	(2.682)	(2.732)	(4.682)	(3.375)
Observations	2,566	2,566	1,303	1,263
Adjusted R2	0.010	0.009	0.013	0.008
Joint test p-value	0.034	0.047	0.012	0.358
Municipality Fixed Effects	Y	Y	Y	Y
Health supply variables	N	Y	Y	Y
Average enrollment rate	0.32		0.36	0.30
Baseline mean	21.8		23.8	16.3

Notes: All regressions include municipality fixed effects and time fixed effects. Each observation is weighted by the population size of the municipality in year 2012. Health supply variables include number of payment centers and health facilities per 1,000 births. We drop observations with stillbirth rates above 300 per 1,000. The sample excludes municipalities with more than 250,000 habitants in year 2012: 6 out of 327 municipalities in the country. Standard errors in parenthesis and are clustered at the municipality level. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Robustness check

The main assumption behind our analysis is that the change in the outcome observed in municipalities with high enrollment rates would have been the same, had they experimented lower levels of enrollment, as those that actually experimented lower levels of enrollment. And vice-versa. In a previous discussion we show that the trends in the rate of stillbirths before year 2009 are uncorrelated with the enrollment rate at the beginning of the BJA. The joint test is provided at the bottom of the Table 3.4, showing a p-value of 0.353. We also estimate the same regression using the enrollment rate for years 2010, 2011 and 2012. We find similar results for these years.

We perform three robustness checks to the main results presented in the paper. Column 2 Table 3.4 presents the results of a falsification test using the future enrollment rate of the program as the treatment variable, showing that there is no significant correlation between these two. This supports, though not sufficiently, the hypothesis that the correlation captured in the principal model is not a spurious one. Columns 3 and 4 of Table 3.4 show that results are robust to dropping observations that show zero stillbirths and to dropping the first year of the data, which shows that the results are not driven by municipalities reporting a stillbirth rate of zero, and that dropping the first year of the data available, believed to be the worst reporting year, does not affect our results.

Finally, we use aggregated 2012 census data on age-cohort population sizes to study the effect of the program on child survival.[19] We use the count in the Census for age cohorts 0 to 5 years in each municipality and a linear regression of the size of each cohort in a municipality as a function of the enrollment rate of pregnant women at the pregnancy stage of the cohort and the enrollment rate of children at the postnatal stage of each cohort. For example, age cohort 1 in year 2012 was exposed to the program during year 2010 in the

19. This analysis is exploratory and we consider it an additional check to the effects on mortality. In particular we check if the effects we are finding on the rate of stillbirths are mirrored by survival of specific age-cohorts. We follow a similar method to Jayachandran (2009).

Table 3.4: Robustness Checks for the Effect of the BJA on the Rate of Stillbirths

	Pre-trend analysis (1)	Falsification test (2)	Dropping zeros (3)	Dropping year 2005 (4)
Enrollment rate in 2009 x Year 2006	0.739			
	(4.476)			
Enrollment rate in 2009 x Year 2007	0.062			
	(5.499)			
Enrollment rate in 2009 x Year 2008	-7.331			
	(5.551)			
Enrollment in t+1		-3.609		
		(2.643)		
Enrollment in t			-3.300	-2.457
			(2.680)	(2.537)
Enrollment in t-1			-6.923**	-6.338**
			(3.077)	(2.731)
Observations	1,291	2,252	1,685	2,238
Adjusted R2	0.004	0.005	0.01	0.01
p - value joint	0.353		0.023	0.025

Notes: Standard errors in parenthesis clustered at the municipality level. (1): Fixed effects at the municipality level and a dummy for each time period are included. Dependent variable is rate of stillbirths. Uses data for years 2005 to year 2008. (2): Fixed effects at the municipality level and a dummy for each time period are included. This regression is identical to (3.1) but uses future enrollment rate as the independent variable. (3): Same as principal model (3.1) excluding observations that reported 0 stillbirths. (4): Same as principal model (3.1) excluding the first year of the data. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

pregnancy stage and during 2011 in the postnatal stage for their first year. For age cohorts 3, 4, and 5 the treatment variables are all equal to zero since they were not exposed to the program at any stage of their life cycle. We find that the enrollment rate of the BJA is positively correlated with the size of the cohort, though only the enrollment rate at the pregnancy stage is significant. For instance a 10% increase in the enrollment of pregnant women is associated with a 0.72% increase in the size of cohorts exposed to the program, as shown in column 1 of Table 3.5 . The relation between cohort size and enrollment rates hold after we control for the sex composition of the cohort and the number of health facilities in

the second column. Furthermore the results show a higher correlation between enrollment rates and cohort size in municipalities where 50% or more households are located in rural areas (Column 3).

Table 3.5: Effect of BJA Intensity on the Size of Age-Cohorts at the Municipality Level

Dep. Var.: Log Size of Age-Cohort	All sample		Rural≥0.5
Prenatal Enrollment	0.075***	0.075***	0.091***
	(0.027)	(0.027)	(0.031)
Postnatal Enrollment	0.004	0.005	0.003
	(0.014)	(0.014)	(0.014)
Observations	1926	1926	1548
Adjusted R2	0.455	0.456	0.466
Municipality Fixed Effects	Y	Y	Y
Health supply variables	N	Y	Y

Notes: The dependent variable is the logarithm of the cohort count. All regressions include municipality fixed effects and time fixed effects. The result is interpreted as the correlation between the rate of enrollment and the percentage change in the size of the cohort. (1): Estimates using municipalities with 200,000 habitants or less. (2): As in (1), including number of health facilities per 1,000 habitants and the ratio of female and male for each cohort. (3): Estimates using municipalities with more than 50% of households in rural areas. Each observation is weighted by the population size of the municipality in year 2012. $*p < 0.10$, $**p < 0.05$, $***p < 0.01$.

The positive association between the implementation of the BJA and the cohort size could be the result of three different acting mechanisms: reduction in infant mortality rates, an increase in the fertility rate, or effects of internal migration. We argue that the results of previously described mortality models using SNIS data support that the main mechanism works through the reduction in rate of stillbirths.

3.5.2 The Effect of the BJA on Prenatal Care Utilization

In the previous section we showed that municipalities with higher rates of BJA enrolment also experienced a steeper reduction in the rate of stillbirths. One of the main mechanisms that could explain this result is that the BJA incentivized women to seek more prenatal care. We estimate regression (3.3) and study the effect of BJA take-up on different outcomes for prenatal care utilization. Table 3.6 presents the results. The first row in column (1) shows that, at the national level, women enrolled in the BJA had their first prenatal check-up 2.5 weeks earlier than women who did not enroll in the program. Relative to the average mean of women not enrolled in the BJA, shown at the bottom of the Panel, the effect on the average week of pregnancy corresponds to a reduction in 17% on the weeks of pregnancy at the first visit. In column (2) we show that this effect holds after we control for a quadratic specification of age at birth, order of pregnancy, and gender of the child. In other words, our results do not change once we account for covariates that proxy risk of the pregnancy or gains in parenting skills. To the extent that these variables are potential sources of bias in our estimates of the effect of the BJA, the results show that using mother fixed effects account for most of the unobserved differences. The next columns show the effect of the BJA for women who live in urban areas and rural areas. Comparing columns (4) to (6), the results show that the effect of the BJA on early initiation of prenatal care is higher in rural areas, where women enrolled had their first prenatal check-up 2.7 weeks earlier than women not enrolled, while in urban areas the effect on the week of pregnancy in the first visit is 1.8 weeks. We also construct the probability that a woman seeks prenatal care for the first time during the first trimester of the pregnancy. Panel B in Table 3.6 shows that women enrolled in the BJA are 8.6 percentage points more likely to have their first check-up during the first trimester of the pregnancy than women not enrolled in the BJA, which corresponds to increasing the probability of early initiation of prenatal care in 11% with respect to the control mean. The marginal effect is higher and only significant for the sub-sample of women

who live in rural areas of the country (see Panel B column (4) and (6)).

Table 3.6: Effect of BJA Enrollment on Utilization of Prenatal Care Services

	All Sample		Urban Households		Rural Households	
	(1)	(2)	(3)	(4)	(5)	(6)
A. Weeks Pregnant At First Prenatal Check-up						
Treatment effect	-2.558***	-2.311***	-2.113**	-1.759**	-2.931***	-2.731***
	(0.542)	(0.546)	(0.816)	(0.848)	(0.716)	(0.687)
Control group mean	13.65		11.92		15.95	
Adjusted R2	0.029	0.034	0.056	0.064	0.033	0.037
B. Probability that First Visit Occurs in the First Trimester						
Treatment effect	0.086***	0.080***	0.064	0.056	0.106***	0.101***
	(0.031)	(0.030)	(0.048)	(0.049)	(0.035)	(0.034)
Control group mean	0.746		0.792		0.687	
Adjusted R2	0.019	0.023	0.024	0.029	0.019	0.022
C. Probability of at Least Four Prenatal Check-ups						
Treatment effect	0.117***	0.103***	0.128**	0.110**	0.110***	0.097***
	(0.028)	(0.028)	(0.049)	(0.047)	(0.026)	(0.025)
Control group mean	0.739		0.807		0.648	
Adjusted R2	0.060	0.065	0.117	0.120	0.040	0.046
D. Probability of Being Attended by Skilled Professional at Birth						
Treatment effect	0.024	0.024	0.000	0.009	0.054**	0.050**
	(0.026)	(0.026)	(0.044)	(0.045)	(0.022)	(0.022)
Control group mean	0.439		0.506		0.351	
Adjusted R2	0.018	0.021	0.034	0.035	0.022	0.024
Observations	5,505	5,505	1,084	1,084	4,421	4,421
Mother fixed effects	Y	Y	Y	Y	Y	Y
Controls for covariates	N	Y	N	Y	N	Y

Notes: This table shows the effect of the BJA enrollment on different outcomes of prenatal care utilization. Columns (1), (3) and (5) report mother fixed-effects regressions estimated with ESNUT 2012 data. Standard errors are in parentheses, clustered at the mother level. Regressions control for mother fixed effects, cohort of birth, a quadratic specification for the mother's age at birth, sex of the child, and ranked order of birth. Observations are weighted using survey weights. *$p < 0.10$, **$p < 0.05$, ***$p < 0.01$.

The next Panel in Table 3.6 shows the effects of the BJA on the probability of completing at least 4 prenatal visits during the pregnancy. The results show that women enrolled in the BJA are 10.3 percentage points more likely to complete at least four prenatal check-ups during their pregnancy than women not enrolled in the BJA, which corresponds to a 16% increase relative to the average rate of 73.9% for women not enrolled. We find no differences between rural and urban areas for this outcome. Finally, we find no significant impacts of the BJA on skilled birth attendance and postpartum care in urban areas (Panel D). In rural areas, however, the program increased the combined probability of having births attended by a professional and postpartum care by 5.0 percentage points, which corresponds to an increase in 14% in the probability of birth and post-partum care relative to the average.

3.5.3 The Effect of the BJA on Children's Utilization of Health Services and Health Outcomes

The BJA was also designed to improve health outcomes of children by incentivizing the number of medical visits for children. Panel B in Figure 3.4 shows the relation between the number of medical visits for children between the 12 and 24 months of age and their birth date. Recall that, at the moment of enrollment, children that were already 12 months of age are not allowed enrollment in the BJA. The graph shows that children born before the eligibility cutoff have on average, approximately 4 visits to the doctor between the 12 and 24 months of age. Children born after the eligibility cutoff have more than five medical visits during the same age. This change is evident at the discontinuity of the age eligibility rule. The ITT estimates in Table 3.7 show that the number of medical visits during this age is 0.824 (CV bandwidth) to 1.08 (CCT bandwidth) higher for children enrolled in the BJA.[20] The LATE estimate shows that the average number of visits among enrolled children ranges

20. To estimate optimal bandwidths we follow recommendations of Calonico et al. (2014). The CCT bandwidth is one that the authors propose while the IK bandwidth corresponds to Imbens and Kalyanaraman (2011) and the CV bandwidth is obtained by a cross-validation method.

from 2.4 (CV bandwidth) to 3.5 visits (CCT bandwidth) higher than those not enrolled. The effect is similar using different bandwidths and specifications of the smoothing function. We find no significant effects on other utilization outcomes such as the probability of having a yellow fever vaccine, MMR vaccine, having a complete immunization or on the probability of consumption of a nutritional supplement administered to children at health clinics.

Table 3.7: Effect of BJA Enrollment on Utilization of Postnatal Care Services. RD estimates

	Number of check-ups at age 12-24 months		Probability of: Yellow fever vaccine	MMR vaccine	Complete immuniz.	Nutritional suppl.
	ITT	LATE	ITT	ITT	ITT	ITT
Non-parametric	1.08	3.546*	0.063	0.018	0.008	0.034
Bandwidth CCT	(0.783)	(2.369)	(0.065)	(0.053)	(0.073)	(0.083)
Observations	413		663	667	740	539
Non-parametric	0.832**	2.737**	0.03	0.025	-0.005	0.06
Bandwidth IK	(0.572)	(1.763)	(0.067)	(0.039)	(0.055)	(0.059)
Observations	1,283		2,356	1,986	1,742	1,656
Non-parametric	0.824**	2.409**	0.027	0.021	0.012	0.015
Bandwidth CV	(0.509)	(1.401)	(0.044)	(0.036)	(0.051)	(0.052)
Observations	1,622		2,334	2,309	2,366	2,473
Semi-parametric	1.123**	3.69***	0.055	0.003	- 0.01	0.07
CV & quadratic poly.	(0.639)	(1.94)	(0.056)	(0.047)	(0.069)	(0.067)
Observations	1,692		2,429	2,400	2,192	2,376

Notes: Standard errors in parenthesis calculated using the method proposed by Calonico, Cattaneo and Titiunik (2014). ITT: Intention to Treat. Differences at the cut-off without adjustments for the probability of being treated. The first stage shows the effect of the eligibility rule on the probability of enrollment into the BJA. LATE: Local Average Treatment Effect is the ITT parameter adjusted by the probability of being treated estimated in the first stage regression. Each row presents the results for different Bandwidths. CV refers to the modified cross-validation procedure used to derive the optimal bandwidth; IK refers to the Imbens-Kalyanarman optimal bandwidth. All regressions are estimated using a local regression with a uniform Kernel weighting method. Standard errors in parenthesis. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

The validity of the results shown in Table 3.7 rely in part on the assumption that at the threshold the assignment to treatment is as good as random. We test for differences at the

Table 3.8: P-value for differences in the eligibility cut-off in RD analysis.

Variables	p-value
Number of check-ups at 0 to 5 months	0.594
Number of check-ups at 0 to 11 months	0.955
Sex, Male=1	0.205
Years of schooling of mother	0.531
Age of mother	0.196
Age of mother at birth	0.201
Household wealth index	0.313
Distance to nearest health facility (km)	0.817
Department, La Paz=1	0.315
Department, Cochabamba=1	0.324
Department, Santa Cruz=1	0.557
Altitude	0.781

Notes: This table shows the p-value of a significance test for the difference in means at the cut-off. The differences in means are obtained by a local linear regression using the bandwidth provided by CCT (2013) for the analysis of our main outcome: number of check-ups at age 12 to 24 months.

cutoff for different covariates. The p-values for whether there are significant differences at the threshold are shown in Table 3.8 . An important feature of the data is that it records age-specific number of medical visits. Children that were eligible to the program in May of 2009 had to be less than one year old, so that comparisons at the cutoff should have only affected the number of visits at months of age close to 12 months rather than number of visits that were not affected by the program, such as visits at age 0 to 5 months. As such we can use the number of medical visits that children had during their first five months as a placebo test. Panel C in Figure 3.4 shows that the average number visits in the first five months of age is smooth around the threshold of enrollment.

We complement the analysis of the effects of the BJA using the fixed effects model explained in section 3.4. Table 3.9 presents the impact of the BJA on the total number of health checkups for children at ages 0 to 23 months. We use the sample of children that are 24 months of age or older. Column (1) in Table 3.9 shows that the estimated effect of the

Table 3.9: Effect of BJA Enrollment on Utilization of Postnatal Care Services

	All Sample		Urban Households		Rural Households	
	(1)	(2)	(3)	(4)	(5)	(6)
A. Number of checkups 0-23 months						
Treatment effect	3.828***	3.619***	4.327***	4.910***	2.844***	2.542***
	(0.748)	(0.739)	(1.373)	(1.477)	(0.794)	(0.783)
Adjusted R2	0.286	0.316	0.416	0.451	0.361	0.377
Observations	1,980		416		1,564	
Control group mean	8.498		8.131		8.992	
B. Probability of yellow fever vaccine						
Treatment effect	0.119***	0.114***	0.147***	0.166***	0.054*	0.051*
	(0.037)	(0.035)	(0.056)	(0.052)	(0.031)	(0.030)
Adjusted R2	0.139	0.158	0.232	0.279	0.091	0.099
Control group mean	0.751		0.707		0.819	
C. Probability of MMR vaccine						
Treatment effect	0.120***	0.117***	0.122***	0.129***	0.080***	0.079***
	(0.031)	(0.031)	(0.046)	(0.046)	(0.025)	(0.025)
Adjusted R2	0.138	0.161	0.225	0.250	0.077	0.079
Control group mean	0.837		0.819		0.864	
D. Probability of complete immunization						
Treatment effect	0.127***	0.124***	0.081*	0.078*	0.138***	0.134***
	(0.031)	(0.030)	(0.042)	(0.043)	(0.034)	(0.033)
Adjusted R2	0.143	0.154	0.296	0.309	0.055	0.060
Control group mean	0.715		0.730		0.692	
E. Probability of taking nutritional supplement						
Treatment effect	0.110**	0.114***	0.162*	0.177**	0.056**	0.060**
	(0.043)	(0.043)	(0.087)	(0.089)	(0.026)	(0.027)
Adjusted R2	0.111	0.117	0.178	0.190	0.100	0.103
Control group mean	0.574		0.486		0.689	
Observations	3,202		624		2,578	
Mother fixed effects	Y	Y	Y	Y	Y	Y
Controls for covariates	N	Y	N	Y	N	Y

Notes: This table shows the effect of the BJA enrollment on different outcomes of post natal care services utilization for children. Columns (1), (3) and (5) report mother fixed-effects regressions estimated with ESNUT 2012 data. Standard errors are in parentheses, clustered at the mother level. Regressions control for a quadratic specification of the mother's age at birth, sex of the child, and dummy variables for birth order and each month of child's age. All observations are weighted using survey weights. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table 3.10: Effect of BJA Enrollment on Health Outcomes of Children

	All Sample		Urban Households		Rural Households	
	(1)	(2)	(3)	(4)	(5)	(6)
A. Z-score length for age (Standard Deviations)						
Treatment effect	0.093	0.089	-0.037	-0.012	0.158	0.157
	(0.110)	(0.109)	(0.181)	(0.192)	(0.135)	(0.134)
Adjusted R2	0.222	0.231	0.380	0.380	0.154	0.157
Observations	2,599		489		2,110	
Control group mean	-1.264		-1.057		-1.539	
B. Probability of chronic malnutrition						
Treatment effect	-0.021	-0.019	-0.018	-0.022	-0.082	-0.081
	(0.053)	(0.053)	(0.081)	(0.082)	(0.055)	(0.055)
Adjusted R2	0.150	0.150	0.249	0.244	0.189	0.193
Observations	2,599		489		2,110	
Control group mean	0.212		0.155		0.283	
C. Probability of Anemia						
Treatment effect	-0.059	-0.057	-0.058	-0.057	-0.061**	-0.058*
	(0.041)	(0.040)	(0.069)	(0.069)	(0.030)	(0.029)
Adjusted R2	0.100	0.106	0.144	0.151	0.093	0.106
Observations	3,820		743		3,077	
Control group mean	0.616		0.523		0.741	
Mother fixed effects	Y	Y	Y	Y	Y	Y
Controls for covariates	N	Y	N	Y	N	Y

Notes: This table shows the effect of the BJA enrollment on different outcomes of health outcomes for children. Columns (1), (3) and (5) report mother fixed-effects regressions estimated with ESNUT 2012 data. Standard errors are in parentheses, clustered at the mother level. Regressions control for a quadratic specification of the mother's age at birth, sex of the child, a quadratic specification of birth interval and dummy variables for birth order and each month of child's age. All observations are weighted using survey weights. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

BJA on the number of checkups is 3.6 additional visits during the child's first two years. In the next column we include a quadratic specification of the mother's age at birth, sex of the child, and dummy variables for birth order and a dummy variable for each month of child's age. The results in column (2) show that the effect of the BJA on number of visits is robust to including controls, which suggests that the fixed effects may already control for most of the potential bias in our estimates. We additionally analyze a set of indicators associated with outcomes of medical visits for children: immunizations and nutritional supplementation. As shown in Table 3.9, BJA has positive impacts on three vaccination indicators: yellow fever vaccine, measles-mumps-rubella (MMR), and an indicator of complete immunization.[21] The program also increased the probability of the consumption of micronutrient supplementation by 11.4 percentage points.

Next we examine the impact of the BJA on child nutritional and growth outcomes measured as height for age z-score (HAZ) in standard deviations, the probability of stunting (HAZ<2) and the probability of anemia. As shown in Table 3.10, we find no statistically significant effects of the BJA on HAZ or on the probability of stunting. However, in rural areas, we find evidence of a 5.8 percentage points decline, significant at the 10% level, in the probability of a child having anemia.

3.6 Discussion

Despite considerable progress in recent years, maternal and child mortality rates as well as child morbidity and stunting continue to be high in Bolivia, especially in poor and vulnerable population segments. Yet many of the most common mortality and morbidity causes are preventable with timely medical care. In Bolivia, as in much of the developing world, pneumonia, diarrhea and other infectious diseases continue to be the main causes of under-five

21. Complete immunization includes the BCG vaccine, third doses of the DPT/Pentavalente vaccine, third doses of the polio vaccine, MMR (SRP) vaccine and the yellow fever vaccine.

mortality, followed by preterm birth and intrapartum-related complications such as perinatal asphyxia. Maternal and perinatal mortality can be effectively reduced with prevention strategies such as prenatal care, delivery and postpartum care and family planning.

In this, paper we study the effects of a CCT program, the Bono Juana Azurduy, designed to increase the demand for basic prenatal care services as well as increase utilization of medical visits by children less than two years old. The results presented here show that the BJA program effectively stimulated the demand for prenatal services nationwide and for the combined package of skilled birth attendance and postpartum care in rural areas. We study the trends in the rate of stillbirths, for years before and after the implementation of the BJA and find that municipalities where the BJA enrolled a higher proportion of women also experienced a steeper decline in the rate of stillbirths. The positive association of mortality at birth and higher enrollment rates of the BJA is a promising signal that CCT models such as the one we study have the potential to improve pregnancy outcomes and the survival and health of newborns.

With respect to child health, the BJA effectively stimulated the utilization of preventive health services, whose protocols include growth and nutrition monitoring, vaccination, and related prevention and counseling strategies. While we find a reduction in the prevalence of anemia in rural areas, we find no evidence of longer-term impacts on children's growth. While final nutrition outcomes such as child growth are influenced by a multi-dimensional set of factors beyond health sector policies alone, sector-specific supply side factors related to the quality of care require further analysis to explain the limited impacts of BJA on final health and nutrition outcomes.

Our cost-effectiveness analysis shows that overall the BJA had a cost of $716.1USD per disability adjusted life year (DALY) averted, making the intervention highly cost-effective when compared to the GDP per capita of $2,480 in 2012. The evidence from this experience suggests that directed monetary incentives paid for critical preventive maternal and child

health services is a promising and cost-effective policy alternative for reducing cultural, economic and behavioral demand side barriers in maternal and child health.

Appendices

APPENDIX A

CHAPTER 2: TEST OF MISREPORTING WEEKS PREGNANT AT 1ST PRENATAL VISIT

One concern is that the financial incentives may cause clinics to misreport the week of pregnancy at the first visit. In this appendix we report the results of test for this behavior. Recall that in our main analysis we construct the week of pregnancy at the first visit using the date of the first visit and the last menstrual date (LMD) as reported by the women. If the latter is not available we use the estimated date of birth (EDD) as recorded by the physician in the first visit. The EDD is calculated off the LMD as reported by the women during her first visit. While clinic medical records should contain both dates, about 10% of records are missing the LMD.

One possible way of misreporting the week of pregnancy at the first visit is to change the LMD and the EDD in the patient's clinical medical record. For instance, if a woman is in her 21st week of pregnancy at the first visit, the physician could add 7 days to the LMD and EDD so that the visit falls into the 20th week of pregnancy. Both would have to be changed in order to deceive the auditors.

To test for this possibility we use gestational age at birth (GAB) in weeks measured by physical examination at the time of birth, registered in the hospital medical record. We then compare the weeks elapsed from the first prenatal visit to the delivery date based on GAB to weeks elapsed from first visit to the delivery date based on EDD. While EDD is collected by the clinic who has an incentive to misreport, the GAB is collected by the hospital at time of delivery where there is no incentive to misreport.

Figure A1 A.1 plots the number of weeks to delivery from the time of the 1st visit based on GAB (y-axis) to the one based on EDD (x-axis). If there is no difference between the two measures, then all of the dates should fall on the 45-degree blue line. There should be some differences as EDD is an estimate that assumes no prematurity at birth, and there could

be data entry in GAB and EDD and recall errors in EDD. Figure A1 shows that almost all of the data embrace the blue 45-degree line and most of the observations off the line are situated above it, consistent with prematurity explaining the differences.

We also explore whether there is any manipulation of the data at the threshold of the 13th week of pregnancy. Figure A2 A.2 shows that there is no discontinuity at this threshold using the test proposed by McCrary (2008) for manipulation at the threshold in studies that use Regression Discontinuity as their research design.

If the clinic changes the EDD in order to capture higher payments, we would expect greater differences, for the treatment group, between GAB and EDD below the 12-week thresholds than above it during the intervention period when the incentives are in force, but no differences in the pre-intervention period. In order to test this, we estimate the following difference in difference regression:

$$W_{ij}^{GAB} = \alpha_j + \beta W_{ij}^{EDD} + \gamma I(W_{ij}^{EDD} < 13) + \delta I(W_{ij}^{EDD} < 13) T_j + \varepsilon_{ij} \tag{A.1}$$

where W_{ij}^{EDD} is weeks of pregnant at the first visit based on EDD for individual i getting care in clinic j, W_{ij}^{GAB} is the number of weeks at the first visit based on GAB for individual i getting care in clinic j, α_j is a clinic fixed effect, $I(W_{ij}^{EDD} < 13)$ is an indicator of whether the clinic reported the first visit to be in the first 12 weeks based on EDD, T_j is an indicator of whether the clinic was actually treated, and ε_{ij} is an error term.

In the absence of misreporting and no prematurity there should be no difference between the two measures and β would have a coefficient of 1. However, because premature births occur before EDD, we expect β to be close to but less than one. Then we can interpret the other coefficients as the effect on $W_{ij}^{GAB} - \beta W_{ij}^{EDD}$ accounting for average weeks of prematurity. So the dependent variable is the error in EDD in forecasting actual delivery date. Equation A.1 takes on a difference in difference interpretation in the sense the we are differencing the change in the forecast error between the pre-intervention and intervention

periods for the group of pregnant women for which a clinic reports as having their first visit before 13 weeks and the group of pregnant women for which a clinic reports having the first visit in week 13 or later. If there is no difference in the error for the treatment group in the post period then δ, the interaction between treatment and reported having the first period before week 13, will be zero. We find no evidence of misclassification by treated clinics (see Table A.1).

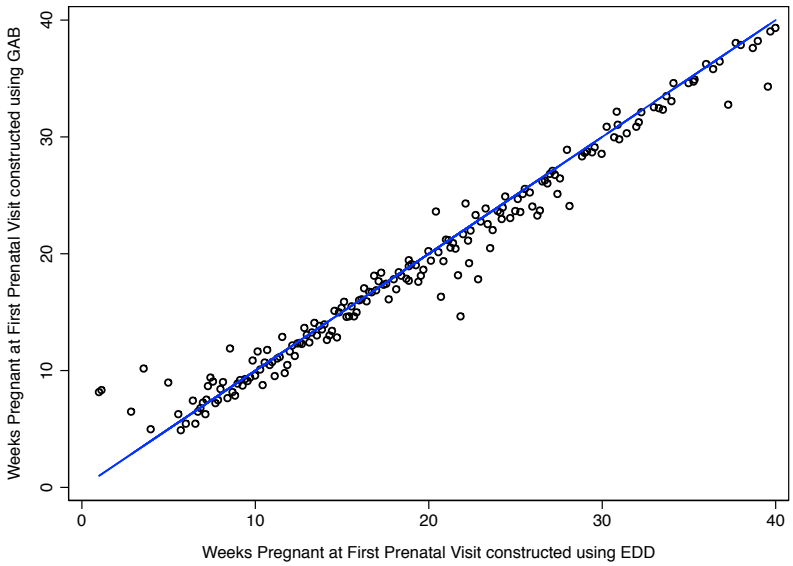

Figure A.1: Comparison of Weeks Pregnant at 1st Prenatal Visit Based on Gestational Age at Birth and Based on Date of Last Menstruation

Notes: Authors' own elaboration based on data from the provincial medical record information system.

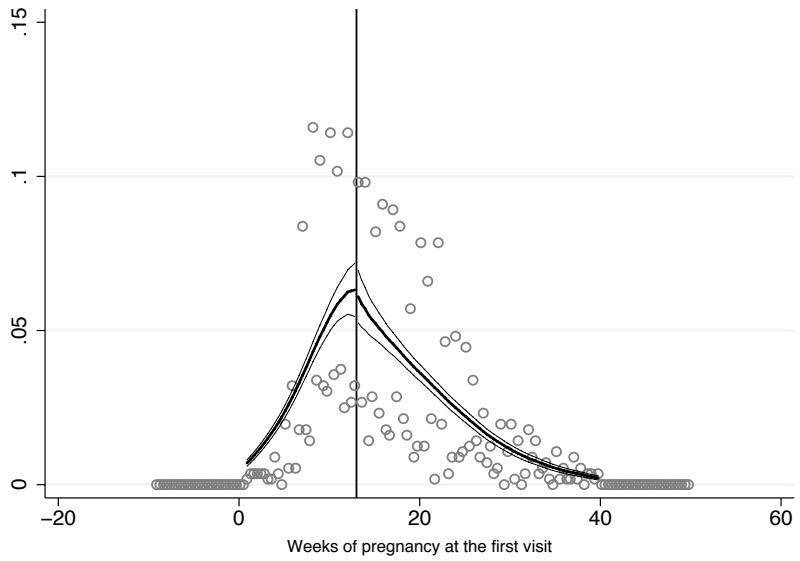

Figure A.2: Test for Misreporting Weeks of Pregnancy at the Threshold of the 13th Week based on the "Manipulation" Test in McCrary (2008)

Notes: Authors' own elaboration based on McCrary (2008).

Table A.1: Test for Misreporting Weeks Pregnant at First Prenatal Visit

Dependent Variable: Weeks Pregnant at 1st Prenatal Visit, by Gestational Age at Birth	
Weeks Pregnant by EDD	0.90***
	(0.02)
I(Weeks Pregnant by EDD<13)	-0.13
	(0.31)
I(Weeks Pregnant by EDD<13) x 1(Treated=1)	-0.03
	(0.44)
Constant	1.33***
	(0.39)
Observations	1730
Adjusted R2	0.82

Notes: The dependent variable is weeks pregnant at the first prenatal visit constructed using gestational age at birth. The independent variable is weeks pregnant at the first visit constructed by using the last day of menstruation or estimated delivery date (EDD). The interaction term interacts a dichotomous indicator for whether the visit was before week 13 and a dichotomous indicator for whether the clinic was actually treated. The regression controls for clinic fixed effects by adding a binary indicator for each clinic in the sample. Standard errors are in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0$

APPENDIX B

CHAPTER 2: ROBUSTNESS TEST RESULTS

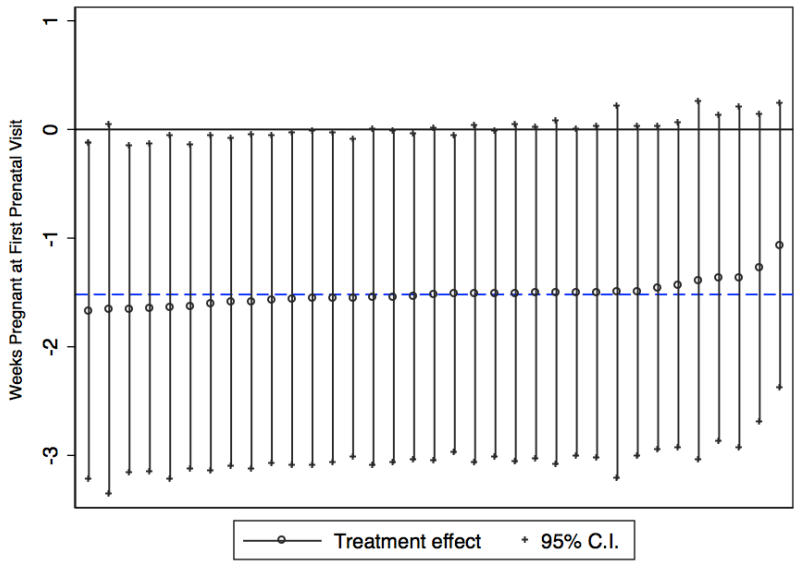

Figure B.1: Estimates of Impact on Weeks Pregnant at First Prenatal Visit Dropping the Observations for each Clinic One at a Time

Notes: This figure plots different treatment effects computed by dropping one clinic at a time for weeks pregnant at the first visit prenatal visit. We run OLS regression of the outcome comparing each clinic assigned to the treatment group to all clinics assigned to the control group pooling the intervention period and post intervention period I (hence May 2010-March 2012). The x-axis is sorted from the lowest to the highest treatment effect. The dashed blue line is the intent-to-treat effect calculated by pooling the intervention and the first post intervention period. The vertical lines are 95% confidence intervals constructed using standard errors obtained from the Wild bootstrap procedure.

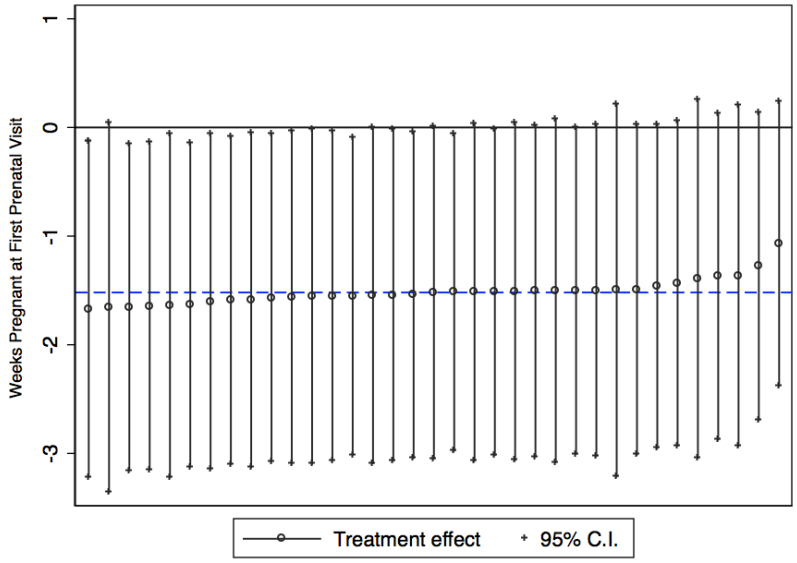

Figure B.2: Estimates of Impact on Weeks First Prenatal Visit Before Week 13 Dropping the Observations for each Clinic One at a Time

Notes: This figure plots different treatment effects computed by dropping one clinic at a time for first prenatal visit before week 13. We run OLS regression of the outcome comparing each clinic assigned to the treatment group to all clinics assigned to the control group pooling the intervention period and post intervention period I (hence May 2010-March 2012). The x-axis is sorted from the lowest to the highest treatment effect. The dashed blue line is the intent-to-treat effect calculated by pooling the intervention and the first post intervention period. The vertical lines are 95% confidence intervals constructed using standard errors obtained from the Wild bootstrap procedure.

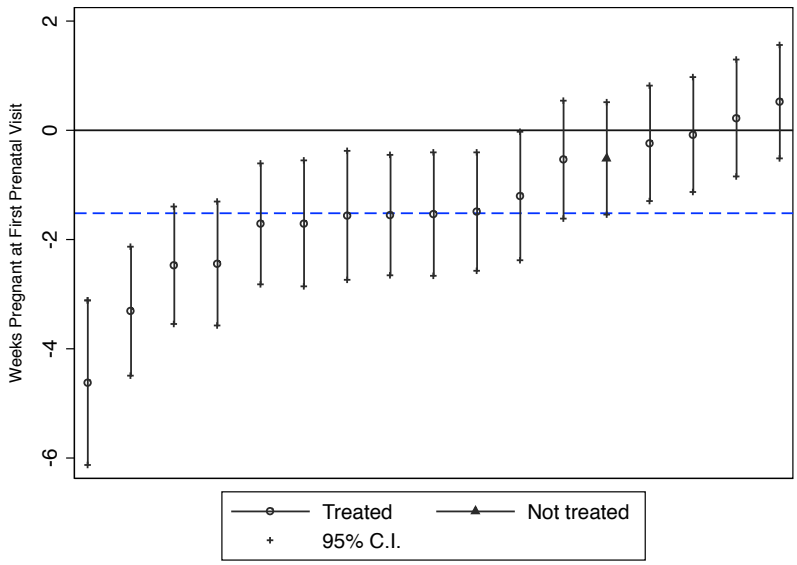

Figure B.3: Individual Clinic Treatment Effects for Weeks Pregnant at First Prenatal Visit

Notes: This figure plots individual clinic treatment effects for the outcome of weeks pregnant at first prenatal visit. We run OLS regression of the outcome comparing each clinic assigned to the treatment group to all clinics assigned to the control group pooling the intervention period and the post-intervention period I (May 2010-March 2012). One treatment clinic is not included because of its insufficient sample size. This clinic corresponds to one of the two that did not take up treatment. The triangle symbol refers to the clinic that was assigned to treatment but did not take up the treatment. The x-axis is sorted from the lowest to the highest clinic-specific impact. The dashed blue line is the intent-to-treat effect calculated by pooling the intervention and the first post intervention period. The vertical lines are 95% confidence intervals constructed using standard errors obtained from the Wild bootstrap procedure.

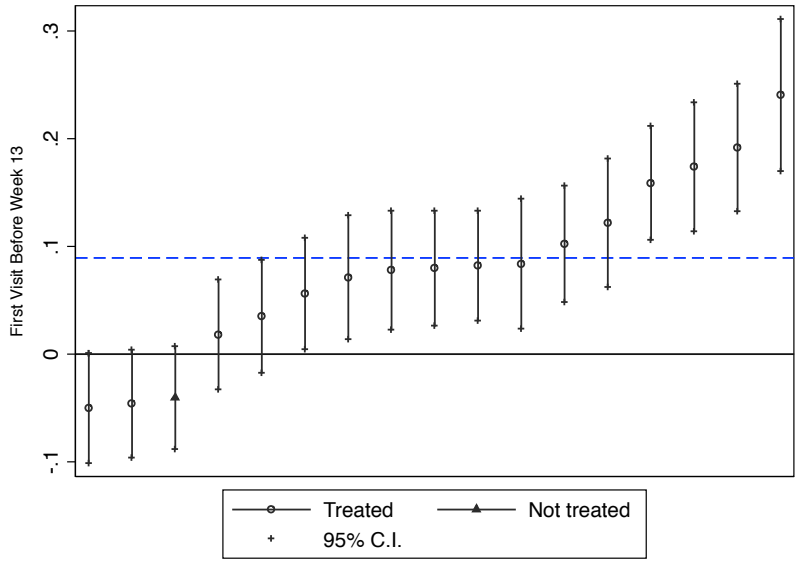

Figure B.4: Individual Clinic Treatment Effects for First Prenatal Visit before Week 13 of Pregnancy

Notes: : This figure plots individual clinic treatment effects for the outcome of first prenatal visit before week 13. We run OLS regression of the outcome comparing each clinic assigned to the treatment group to all clinics assigned to the control group pooling the intervention period and post intervention period I (hence May 2010-March 2012). One treatment clinic is not included because of its insufficient sample size. This clinic corresponds to one of the two that did not take up treatment. The triangle symbol refers to the clinic that was assigned to treatment but did not take up the treatment. The x-axis is sorted from the lowest to the highest clinic-specific impact. The dashed blue line is the intent-to-treat effect calculated by pooling the intervention and the first post intervention period. The vertical lines are 95% confidence intervals constructed using standard errors obtained from the Wild bootstrap procedure.

Table B.1: Robustness Tests for Weeks Pregnant at First Prenatal Visit

	(1)	(2)	(3)
	Intervention Period	Post-Intervention Period I	Post-Intervention Period II
A. Results from Table 2.4			
Treatment	-1.47**	-1.63**	-2.47**
	(0.71)	(0.75)	(1.02)
Large Sample p-value	0.04	0.03	0.02
Wild Bootstrapped p-value	0.08	0.03	0.03
Control Group Mean	17.80	17.90	20.10
Sample Size	769	1.296	710
B. Estimates Using Restricted Sample			
Treatment	-1.47*	-2.01***	-2.01*
	(0.77)	(0.70)	(1.11)
Large Sample p-value	0.06	0.00	0.07
Wild Bootstrapped p-value	0.09	0.02	0.12
Control Group Mean	17.96	18.32	17.01
Sample Size	760	1.326	425
C. Difference-in-Differences Estimates			
Treatment	-1.35**	-1.74***	-2.35*
	(0.64)	(0.63)	(1.31)
Large Sample p-value	0.036	0.005	0.072
Wild Bootstrapped p-value	0.060	0.014	0.144
Control Group Mean	17.80	17.90	20.10
Sample Size	4.015	4.015	4.015

Notes: This table reports LATE estimates of the treatment effect of the modified fee schedule on weeks pregnant at 1st prenatal visit. The p-values are for 2-sided hypothesis tests of the null that the difference is equal to zero. We present both the p-value computed for large samples and a Wild bootstrapped p-value that is robust in samples with small numbers of clusters (Cameron et al., 2008). Our Wild bootstrap procedure assigns symmetric weights and equal probability after re-sampling residuals (Davidson and Flachaire, 2008) and uses 999 replications. Column (1) reports the results for the sample observed in an 8-month intervention period (May 2010 - December 2010). Column (2) reports the results for the sample observed in the 15-month period following the end of the intervention (January 2011 - March 2012). Column (3) reports the results for the 9-month period after the change in the coding of the first prenatal visit (April 2012 - December 2012). Standard errors are in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0$

Table B.2: Robustness Tests for First Prenatal Visit before Week 13

	(1)	(2)	(3)
	Intervention Period	Post-Intervention Period I	Post-Intervention Period II
A. Results from Table 2.4			
Treatment	0.11**	0.08**	0.08**
	(0.04)	(0.04)	(0.04)
Large Sample p-value	0.01	0.02	0.04
Wild Bootstrapped p-value	0.03	0.05	0.06
Control Group Mean	0.31	0.34	0.27
Sample Size	769	1.296	710
B. Estimates Using Restricted Sample			
Treatment	0.09**	0.10**	0.10*
	(0.04)	(0.04)	(0.06)
Large Sample p-value	0.03	0.01	0.08
Wild Bootstrapped p-value	0.08	0.02	0.11
Control Group Mean	0.31	0.33	0.36
Sample Size	760	1.326	425
C. Difference-in-Differences Estimates			
Treatment	0.09*	0.07	0.07
	(0.05)	(0.05)	(0.06)
Large Sample p-value	0.08	0.11	0.23
Wild Bootstrapped p-value	0.13	0.17	0.24
Control Group Mean	0.31	0.34	0.27
Sample Size	4.015	4.015	4.015

Notes: This table reports LATE estimates of the treatment effect of the modified fee schedule an indicator of whether the 1st prenatal visit occurred before week 13 of pregnancy. The p-values are for 2-sided hypothesis tests of the null that the difference is equal to zero. We present both the p-value computed for large samples and a Wild bootstrapped p-value that is robust in samples with small numbers of clusters (Cameron et al., 2008). Our Wild bootstrap procedure assigns symmetric weights and equal probability after re-sampling residuals (Davidson and Flachaire, 2008) and uses 999 replications. Column (1) reports the results for the sample observed in an 8-month intervention period (May 2010 - December 2010). Column (2) reports the results for the sample observed in the 15-month period following the end of the intervention (January 2011 - March 2012). Column (3) reports the results for the 9-month period after the change in the coding of the first prenatal visit (April 2012 - December 2012). Standard errors are in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0$

APPENDIX C

CHAPTER 2: ITT RESULTS

Table C.1: ITT Estimates of the Effect of Temporary Incentives on Timing of First Prenatal Visit

	(1)	(2)	(3)
	Intervention Period	Post-Intervention Period I	Post-Intervention Period II
A. Weeks Pregnant at 1st Prenatal Visit			
Treatment	-1.39**	-1.59**	-2.47**
	(0.67)	(0.73)	(1.02)
Large Sample p-value	0.04	0.03	0.02
Wild Bootstrapped p-value	0.09	0.03	0.03
Control Group Mean	17.80	17.90	20.10
Sample Size	769	1.296	710
B. First Prenatal Visit Before Week 13 of Pregnancy			
Treatment	0.10***	0.08**	0.08**
	(0.04)	(0.04)	(0.04)
Large Sample p-value	0.01	0.02	0.04
Wild Bootstrapped p-value	0.03	0.05	0.08
Control Group Mean	0.31	0.34	0.27
Sample Size	769	1.269	710

Notes: This table reports ITT estimates of the treatment effect of the modified fee schedule on indicators of the timing of the 1st prenatal visit. The LATE estimates are reported in Table 2.4. The differences are estimated from OLS regressions of the dependent variable on an indicator for clinic treatment random assignment. The p-values are for 2-sided hypothesis tests of the null that the difference is equal to zero. We present both the p-value computed for large samples and a Wild bootstrapped p-value that is robust in samples with small numbers of clusters (Cameron et al., 2008). Our Wild bootstrap procedure assigns symmetric weights and equal probability after re-sampling residuals (Davidson and Flachaire, 2008) and uses 999 replications. Column (1) reports the results for the sample observed in an 8-month intervention period (May 2010 - December 2010). Column (2) reports the results for the sample observed in the 15-month period following the end of the intervention (January 2011 - March 2012). Column (3) reports the results for the 9-month period after the change in the coding of the first prenatal visit (April 2012 - December 2012). Standard errors are in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0$

Table C.2: ITT of Cross-Price Effects (Spillover)

	(1)	(2)
	Intervention Period	Post-Intervention Period I
A. Tetanus Vaccine		
Treatment	0.02	-0.02
	(0.07)	(0.05)
Large Sample p-value	0.76	0.62
Wild Bootstrapped p-value	0.80	0.59
Control Group Mean	0.79	0.84
Sample Size	769	1.053
A. Number of visits		
Treatment	0.37	0.50
	(0.32)	(0.57)
Large Sample p-value	0.24	0.38
Wild Bootstrapped p-value	0.27	0.40
Control Group Mean	4.05	4.40
Sample Size	769	1.053

Notes: This table reports ITT estimates of the treatment effect of the modified fee schedule on indicators of other services. The LATE estimates are reported in Table 2.5. The differences are estimated from OLS regressions of the dependent variable on an indicator for clinic treatment random assignment. The p-values are for 2-sided hypothesis tests of the null that the difference is equal to zero. We present both the p-value computed for large samples and a Wild bootstrapped p-value that is robust in samples with small numbers of clusters (Cameron et al., 2008). Our Wild bootstrap procedure assigns symmetric weights and equal probability after re-sampling residuals (Davidson and Flachaire, 2008) and uses 999 replications. Column (1) reports the results for the sample observed in an 8-month intervention period (May 2010 - December 2010). Column (2) reports the results for the sample observed in the 15-month period following the end of the intervention (January 2011 - March 2012). Standard errors are in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0$

Table C.3: ITT Effects of Incentives on Birth Outcomes

	(1)	(2)
	Intervention Period	Post-Intervention Period I
A. Birth Weight		
Treatment	-34.88	24.48
	(45.38)	(39.63)
Large Sample p-value	0.44	0.54
Wild Bootstrapped p-value	0.46	0.57
Control Group Mean	3304.82	3279.13
Sample Size	555	802
B. Low Birth Weight		
Treatment	0.01	-0.01
	(0.02)	(0.01)
Large Sample p-value	0.63	0.60
Wild Bootstrapped p-value	0.61	0.63
Control Group Mean	0.05	0.06
Sample Size	555	802
B. Premature		
Treatment	0.03	-0.04*
	(0.03)	(0.02)
Large Sample p-value	0.31	0.08
Wild Bootstrapped p-value	0.32	0.09
Control Group Mean	0.09	0.12
Sample Size	414	708

Notes: This table reports ITT estimates of the treatment effect of the modified fee schedule for on indicators of birth outcomes. The LATE estimates are reported in Table 2.6. The observations include woman for whom we are able to obtain information on birth outcomes provided in public hospital birth records. The differences are estimated from OLS regressions of the dependent variable on an indicator for clinic treatment random assignment. The p-values are for 2-sided hypothesis tests of the null that the difference is equal to zero. We present both the p-value computed for large samples and a Wild bootstrapped p-value that is robust in samples with small numbers of clusters (Cameron et al., 2008). Our Wild bootstrap procedure assigns symmetric weights and equal probability after re-sampling residuals (Davidson and Flachaire, 2008) and uses 999 replications. Column (1) reports the results for the sample observed in an 8-month intervention period (May 2010 - December 2010). Column (2) reports the results for the sample observed in the 15-month period following the end of the intervention (January 2011 - March 2012). Standard errors are in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0$

APPENDIX D

CHAPTER 2: SURVEY OF CLINIC MEDICAL DIRECTORS

In collaboration with the Provincial Management Unit of the program (UGPS), we conducted a short of clinics that participated in the pilot. The survey aimed to measure the absolute and relative importance of seven different prenatal care procedures including initiating prenatal care prior to week 13 of pregnancy. The absolute scores range from 1 to 5, with 5 being the highest score in terms of importance, and an additional option of zero indicating that the procedure is not appropriate for a pregnant woman. Hence, the absolute score ranges from 0 to 5 points. The relative ranking aimed to sort the seven practices from 1 to 7, with 1 being the highest ranking. In practice however, the survey instrument allowed the respondent to repeat numbers.

The survey was sent out to by email to clinics directors (or the next person in rank). Sixty-seven percent of the clinics responded to the survey. Appendix D Table D1 D.1 shows that there are no significant differences in baseline characteristics between clinics that responded to the survey and clinics that did not respond. In addition, we account for survey non-response using Inverse Probability Weighting based on the logistic regression reported in Table D2 D.2 (Wooldridge, 2007). We report results for both IPW and non-IPW regressions.

Figures 2.6 do not suggest any difference in the absolute score and relative ranking of the procedures between treatment and control clinics. To test for the significance of the differences between the two groups, we run an OLS regression of the absolute score and the relative ranking against a binary indicator for treatment. To account for the small sample size we also compute the p-value for the differences in means permuting our data and using a random sample of 10,000 permutations. The results are shown in Table D.3.

Survey Questionnaire

We ask for your collaboration in completing a brief survey about prenatal care services

provided at your health facility.

Important: When answering the survey, please think of a hypothetical case of a woman with the following characteristics:

- 25 years old
- Living in the same neighborhood where your health facility is located
- Without any apparent sign of disease
- 6 weeks pregnant
- Had a previous low-risk pregnancy

1. *Please assign a score between 1 to 5 to each of the following services that could be delivered to the pregnant woman presented in the hypothetical case.*

- **1** corresponds to a service to which you assign the lowest importance
- **5** corresponds to a service to which you assign the highest importance

	1	**2**	**3**	4	5	Not appropriate for a pregnant woman
Prenatal ultrasound	○	○	○	○	○	○
Thorax X-Ray	○	○	○	○	○	○
First prenatal visit before week 13 of pregnancy	○	○	○	○	○	○
Bio-psycho-social pregnancy counseling visit	○	○	○	○	○	○
Combined Diphtheria/Tetanus vaccine	○	○	○	○	○	○
Blood test with serology	○	○	○	○	○	○
Blood test without serology	○	○	○	○	○	○

2. *Please rank in order of priority (from 1 to 7) the following 7 health services that could be delivered to the pregnant woman of the hypothetical case.*

- **1** corresponds to the service you would prioritize the most

- **7** corresponds to the service you would prioritize the least

Prenatal ultrasound	
Thorax X-Ray	
First prenatal visit before week 13 of pregnancy	
Bio-psycho-social pregnancy counseling visit	
Combined Diphtheria/Tetanus vaccine	
Blood test with serology	
Blood test without serology	

Table D.1: Baseline Characteristics of Clinics, by Online Survey Response Status

	Non-respondent	Respondent	P-value	Obs.
Number of Pregnant Women Attended per Year	48.60	54.90	0.33	36
Weeks Pregnant at 1st Prenatal Visit	17.04	16.77	0.15	36
1st Visit before Week 13 of Pregnancy	0.34	0.36	0.27	36
% of Pregnant Women who are Plan Nacer Beneficiaries	0.61	0.64	0.59	36
Tetanus Vaccine During Prenatal Visit	0.76	0.81	0.22	36
Number of Prenatal Visits	4.26	4.42	0.72	36
Birth Weight (Grams)	3,283	3,320	0.33	36
Gestational Age (Weeks)	38.65	38.47	0.57	31
Low Birth Weight ($<$ 2500 Grams)	0.06	0.07	0.73	31
Premature (Gestational Age $<$ 37 Weeks)	0.10	0.12	0.60	31

Notes: This table reports the means of baseline characteristics for clinics that responded to the May 2015 online survey and for clinics that did not respond. The characteristics are taken from the medical records information system (2009). The p-values for the tests of differences in means are computed using permutation tests that are robust for small sample sizes.

Table D.2: Probability of Responding to the Online Survey, Logit Coefficients and Marginal Effects

	Coefficient	Marg. Eff.
Treatment Group	1.498	0.274
	(1.111)	(0.180)
Birth Weight (grams)	0.100	0.018
	(1.076)	(0.196)
Weeks Pregnant at 1st Prenatal Visit	-0.594	-0.109
	(0.648)	(0.121)
1st Visit before Week 13 of Pregnancy	-3.590	-0.657
	(9.026)	(1.670)
% of Pregnant Women who are Plan Nacer Beneficiaries	1.620	0.296
	(4.359)	(0.774)
Tetanus Vaccine During Prenatal Visit	3.350	0.613
	(3.817)	(0.646)
Number of Prenatal Visits	-0.099	-0.018
	(0.559)	(0.101)
Constant	7.644	
	(18.248)	
Observations	36	36

Notes: This table reports the coefficients and marginal effects from a Logit regression that estimates the probability that a clinic responded to the May 2015 online survey.

Table D.3: Differences in Absolute Score and Relative Ranking of Early Prenatal Care

	Absolute Score		Relative Ranking	
	(1)	(2)	(3)	(4)
	OLS	OLS-IPW	OLS	OLS-IPW
Difference (Treatment - Control)	0.20	0.13	0.10	0.14
	(0.22)	(0.92)	(0.21)	(0.89)
Large Sample p-value	0.38	0.89	0.65	0.88
Permutation p-value	0.35	1.00	0.46	0.99
Observations	20	20	20	20
Control group mean	4.57	1.88	4.66	1.88

Notes: Column (1) shows the differences between treatment and control clinics in the absolute score assigned to the practice of early prenatal care without any adjustment of sample loss. Column (2) adjusts for sample loss by Inverse Probability Weighting. Column (3) shows the differences between treatment and control clinics in the relative ranking assigned to early prenatal care among seven different practices. Column (4) is the same as Column (3) but adjusts for sample loss by Inverse Probability Weighting. (Wooldridge, 2007). The coefficients are obtained from an OLS regression of each outcome against a treatment binary indicator. The third row shows the P-value obtained from permuting the data using a random sample of 10,000 permutations. Standard errors are in parentheses. We lose one observation in each case because of missing data in each specific question

REFERENCES

Acland, D. and M. R. Levy (2015). Naiveté, projection bias, and habit formation in gym attendance. *Management Science 61*(1), 146–160.

Almond, D., H. W. Hoynes, and D. W. Schanzenbach (2011). Inside the war on poverty: The impact of food stamps on birth outcomes. *The Review of Economics and Statistics 93*(2), 387–403.

Alwin, D. F. (2007). *Margins of error: A study of reliability in survey measurement*, Volume 547. John Wiley & Sons.

Aquilino, W. S. and L. A. L. Sciuto (1990). Effects of interview mode on self-reported drug use. *Public Opinion Quarterly 54*(3), 362–393.

Ashraf, N., O. Bandiera, S. S. Lee, et al. (2014). Do-gooders and go-getters: career incentives, selection, and performance in public service delivery. *STICERD-Economic Organisation and Public Policy Discussion Papers Series 54*.

Atkin, D., A. Chaudhry, S. Chaudry, A. K. Khandelwal, and E. Verhoogen (2015). Organizational barriers to technology adoption: Evidence from soccer-ball producers in pakistan. Technical report, National Bureau of Economic Research.

Baker, G. P., M. C. Jensen, and K. J. Murphy (1988). Compensation and incentives: Practice vs. theory. *The journal of Finance 43*(3), 593–616.

Baker, L. C. (2001). Managed care and technology adoption in health care: evidence from magnetic resonance imaging. *Journal of health economics 20*(3), 395–421.

Baker, L. C. and C. S. Phibbs (2002). Managed care, technology adoption, and health care: the adoption of neonatal intensive care. *The RAND Journal of Economic 33*(3), 524–548.

Banerjee, A. V. and E. Duflo (2011). *Poor Economics: Barefoot Hedge-fund Managers, DIY Doctors and the Surprising Truth about Life on Less Than 1 [dollar] a Day*. Penguin Books.

Barber, S. L. and P. J. Gertler (2009). Empowering women to obtain high quality care: evidence from an evaluation of mexico's conditional cash transfer programme. *Health Policy and Planning 24*(1), 18–25.

Barham, T. (2011). A healthier start: the effect of conditional cash transfers on neonatal and infant mortality in rural mexico. *Journal of Development Economics 94*(1), 74–85.

Basinga, P., P. J. Gertler, A. Binagwaho, A. L. Soucat, J. Sturdy, and C. M. Vermeersch (2011). Effect on maternal and child health services in rwanda of payment to primary health-care providers for performance: an impact evaluation. *The Lancet 377*(9775), 1421–1428.

Bech, M., T. Christiansen, K. Dunham, J. Lauridsen, C. H. Lyttkens, K. McDonald, and A. McGuire (2009). The influence of economic incentives and regulatory factors on the adoption of treatment technologies: a case study of technologies used to treat heart attacks. *Health economics 18*(10), 1114–1132.

Belli, R. F., M. W. Traugott, and M. 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), 479–498.

Berwick, D. M. (2003). Disseminating innovations in health care. *Jama 289*(15), 1969–1975.

Besley, T. and S. Coate (1992). Understanding welfare stigma: Taxpayer resentment and statistical discrimination. *Journal of Public Economics 48*(2), 165–183.

Biemer, P. P., R. M. Groves, L. E. Lyberg, N. A. Mathiowetz, and S. Sudman (2011). *Measurement errors in surveys*. John Wiley & Sons.

Black, D., S. Sanders, and L. Taylor (2003). Measurement of higher education in the census and current population survey. *Journal of the American Statistical Association 98*(463), 545–54.

Black, D. A., M. C. Berger, and F. A. Scott (2000). Bounding parameter estimates with nonclassical measurement error. *Journal of the American Statistical Association 95*(451), 739–748.

Blair, J., G. Menon, and B. Bickart (2004). Measurement Effects in Self vs. Proxy Response to Survey Questions: An Information-Processing Perspective. In P. P. Biemer, R. M. Groves, L. E. Lyberg, N. A.thiowetz, and S. Sudman (Eds.), *Measurement Errors in Surveys*, pp. 145–166. John Wiley & Sons, Inc.

Blank, R. M. (2002). Evaluating welfare reform in the united states. *Journal of Economic Literature 40*(4), 1105–66.

Blank, R. M. and P. 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.

Blattberg, R. C. and S. A. Neslin (1990). *Sales promotion: Concepts, methods, and strategies*. Prentice Hall Englewood Cliffs, NJ.

Bloom, N., B. Eifert, A. Mahajan, D. McKenzie, and J. Roberts (2012). Does management matter? evidence from india. *The Quarterly Journal of Ecoomics 128*(1), 1–51.

Bollinger, C. R. (1996). Bounding mean regressions when a binary regressor is mismeasured. *Journal of Econometrics 73*(2), 387–399.

Bollinger, C. R. and M. H. David (1997). Modeling discrete choice with response error: Food stamp participation. *Journal of the American Statistical Association 92*(439), 827–835.

Bollinger, C. R. and M. H. David (2001). Estimation with response error and nonresponse: food-stamp participation in the sipp. *Journal of Business & Economic Statistics 19*(2), 129–141.

Bollinger, C. R. and B. T. Hirsch (2006). Match bias from earnings imputation in the current population survey: The case of imperfect matching. *Journal of Labor Economics 24*(3), 483–519.

Bonfrer, I., R. Soeters, E. van de Poel, O. Basenya, G. Longin, F. van de Looij, and E. van Doorslaer (2013). The effects of performance-based financing on the use and quality of health care in burundi: an impact evaluation. *The Lancet 381*, S19.

Bordalo, P., N. Gennaioli, and A. Shleifer (2012). Salience theory of choice under risk. *The Quarterly Journal of Ecoomics 127*(3), 1243–1285.

Bordalo, P., N. Gennaioli, and A. Shleifer (2013). Salience and consumer choice. *Journal of Political Economy 121*(5), 803–843.

Bound, J., C. Brown, and N. Mathiowetz (2001). Measurement error in survey data. *Handbook of Econometrics 5*, 3705–3843.

Bradburn, N. M. (2016). Surveys as social interactions. *Journal of Survey Statistics and Methodology*, smv037.

Bresnahan, T. F. and M. Trajtenberg (1995). General purpose technologies? engines of growth? *Journal of econometrics 65*(1), 83–108.

Brittingham, A., R. Tourangeau, and W. 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, K., G. Müller, and R. T. Riphahn (2014). Who misreports welfare receipt in surveys? *Applied Economics Letters 21*(12), 812–816.

Bruckmeier, K., G. Müller, and R. T. Riphahn (2015). Survey misreporting of welfare receipt?respondent, interviewer, and interview characteristics. *Economics Letters 129*, 103–107.

Butler, J. S., R. V. Burkhauser, J. M. Mitchell, and T. P. Pincus (1987). Measurement error in self-reported health variables. *The Review of Economics and Statistics*, 644–650.

Cabana, M. D., C. S. Rand, N. R. Powe, A. W. Wu, M. H. Wilson, P.-A. C. Abboud, and H. R. Rubin (1999). Why don't physicians follow clinical practice guidelines? a framework for improvement. *Jama 282*(15), 1458–1465.

Calonico, S., M. D. Cattaneo, R. Titiunik, et al. (2014). Robust data-driven inference in the regression-discontinuity design. *Stata Journal 14*(4), 909–946.

Cameron, A. C., J. B. Gelbach, and D. L. Miller (2008). Bootstrap-based improvements for inference with clustered errors. *The Review of Economics and Statistics 90*(3), 414–427.

Campbell, O. M. and W. J. Graham (2006a). Strategies for reducing maternal mortality: getting on with what works. *The lancet 368*(9543), 1284–1299.

Campbell, O. M. and W. J. Graham (2006b). Strategies for reducing maternal mortality: getting on with what works. *The lancet 368*(9543), 1284–1299.

Campbell, S., D. Reeves, E. Kontopantelis, E. Middleton, B. Sibbald, and M. Roland (2007). Quality of primary care in england with the introduction of pay for performance. *New England Journal of Medicine 357*(2), 181–190.

Carroli, G., C. Rooney, and J. Villar (2001). How effective is antenatal care in preventing maternal mortality and serious morbidity? an overview of the evidence. *Paediatric and perinatal Epidemiology 15*(s1), 1–42.

Carroli, G., J. Villar, G. Piaggio, D. Khan-Neelofur, M. Gülmezoglu, M. Mugford, P. Lumbiganon, U. Farnot, P. Bersgjø, W. A. C. T. R. Group, et al. (2001). Who systematic review of randomised controlled trials of routine antenatal care. *The Lancet 357*(9268), 1565–1570.

Carroli, G., J. Villar, G. Piaggio, D. Khan-Neelofur, M. Glmezoglu, M. Mugford, P. Lumbiganon, U. Farnot, P. Bersgj, W. A. C. T. R. Group, and others (2001). WHO systematic review of randomised controlled trials of routine antenatal care. *The Lancet 357*(9268), 1565–1570.

Carroll, G. R. and M. T. Hannan (2000). *The demography of corporations and industries*. Princeton University Press.

Cartwright, A. (1957). The effect of obtaining information from different informats on a family morbidity inquiry. *Applied Statistics*, 18–25.

Cawley, J. and J. A. Price (2013). A case study of a workplace wellness program that offers financial incentives for weight loss. *Journal of health economics 32*(5), 794–803.

Cecchini, S. and A. Madariaga (2011). Conditional cash transfer programmes: the recent experience in latin america and the caribbean. *Cuadernos de la CEPAL* (95).

Cecchini, S. and F. V. Soares (2015). Conditional cash transfers and health in latin america. *The Lancet 385*(9975), e32–e34.

Celhay, P., B. Meyer, and N. Mittag (2015). Measurement Error in Program Participation.

Celhay, P. A., B. D. Meyer, and N. Mittag (2016). Measurement error in program participation. Technical report, Harris School of Public Policy, University of Chicago.

Charness, G. and U. Gneezy (2009). Incentives to exercise. *Econometrica 77*(3), 909–931.

Clemens, J. and J. D. Gottlieb (2014). Do physicians' financial incentives affect medical treatment and patient health? *The American economic review 104*(4), 1320.

Coffman, K. B., L. C. Coffman, and K. M. M. Ericson (2013). The size of the lgbt population and the magnitude of anti-gay sentiment are substantially underestimated. Technical report, National Bureau of Economic Research.

Coleman, J. S., E. Katz, H. Menzel, et al. (1966). *Medical innovation: A diffusion study*. Bobbs-Merrill Indianapolis.

Comin, D. and B. Hobiijn (2010). An exploration of technology diffusion. *The American economic review 100*(5), 2031–59.

Conley, T. G. and C. R. Udry (2010). Learning about a new technology: Pineapple in ghana. *The American Economic Review 100*(1), 35–69.

Currie, J., J. Grogger, G. Burtless, and R. F. Schoeni (2001). Explaining recent declines in food stamp program participation [with comments]. *Brookings-Wharton papers on urban affairs*, 203–244.

Cutler, D. M. (2007). The lifetime costs and benefits of medical technology. *Journal of Health Economics 26*(6), 1081–1100.

Cutler, D. M. and R. S. Huckman (2003). Technological development and medical productivity: the diffusion of angioplasty in new york state. *Journal of Health Economics 22*(2), 187–217.

Danielson, C. and J. A. Klerman (2006). Why did the food stamp caseload decline (and rise)?

Das, J. and P. J. Gertler (2007). Variations in practice quality in five low-income countries: a conceptual overview. *Health Affairs 26*(3), w296–w309.

Das, J. and J. Hammer (2005). Which doctor? combining vignettes and item response to measure clinical competence. *Journal of Development Economics 78*(2), 348–383.

Das, J., J. Hammer, and K. Leonard (2008). The quality of medical advice in low-income countries. *The Journal of Economic Perspectives 22*(2), 93–114.

David, M. (1962). The validity of income reported by a sample of families who received welfare assistance during 1959. *Journal of the American Statistical Association 57*(299), 680–685.

David, M. H. and C. R. Bollinger (2000). Differential reporting of food stamps and afdc" explanations and conjectures.

David, P. A. (1990). The dynamo and the computer: an historical perspective on the modern productivity paradox. *The American Economic Review 80*(2), 355–361.

Davidson, R. and E. Flachaire (2008). The wild bootstrap, tamed at last. *Journal of Econometrics 146*(1), 162–169.

De Brauw, A., D. O. Gilligan, J. Hoddinott, and S. Roy (2014). The impact of bolsa família on women?s decision-making power. *World Development 59*, 487–504.

de Leeuw, E. D. (1992). *Data quality in mail, telephone and face to face surveys.* ERIC.

De Mel, S., C. McIntosh, and C. Woodruff (2013). Deposit collecting: Unbundling the role of frequency, salience, and habit formation in generating savings. *The American Economic Review 103*(3), 387–392.

De Walque, D., P. J. Gertler, S. Bautista-Arredondo, A. Kwan, C. Vermeersch, J. de Dieu Bizimana, A. Binagwaho, and J. Condo (2015). Using provider performance incentives to increase hiv testing and counseling services in rwanda. *Journal of health economics 40*(1), 1–9.

Deaton, A. (1997). *The analysis of household surveys: a microeconometric approach to development policy*. World Bank Publications.

DellaVigna, S. (2009). Psychology and economics: Evidence from the field. *Journal of Economic Literature 47*(2), 315–372.

DeMaio, T. J. (1984). Social desirability and survey. *Surveying Subjective Phenomena,— 2*, 257.

Duflo, E., M. Kremer, and J. Robinson (2011). Nudging farmers to use fertilizer: Theory and experimental evidence from kenya. *The American Economic Review 101*(6), 2350–90.

Duncan, G. J. and D. H. Hill (1985). An investigation of the extent and consequences of measurement error in labor-economic survey data. *Journal of Labor Economics*, 508–532.

Dupas, P. (2014). Short-run subsidies and long-run adoption of new health products: Experimental evidence from kenya. *Econometrica 82*, 197–228.

Eisenhower, D., N. A. Mathiowetz, and D. Morganstein (2004). Recall Error: Sources and Bias Reduction Techniques. In P. P. Biemer, R. M. Groves, L. E. Lyberg, N. A.thiowetz, and S. Sudman (Eds.), *Measurement Errors in Surveys*, pp. 125–144. John Wiley & Sons, Inc.

Ensor, T. and S. Cooper (2004). Overcoming barriers to health service access: influencing the demand side. *Health policy and planning 19*(2), 69–79.

Essig, L. and J. K. Winter (2009). Item non-response to financial questions in household surveys: An experimental study of interviewer and mode effects*. *Fiscal Studies 30*(3-4), 367–390.

Fernald, L. C., P. J. Gertler, and L. M. Neufeld (2008). Role of cash in conditional cash transfer programmes for child health, growth, and development: an analysis of mexico's oportunidades. *The Lancet 371*(9615), 828–837.

Fiszbein, A., N. R. Schady, and F. H. Ferreira (2009). *Conditional cash transfers: reducing present and future poverty*. World Bank Publications.

Flores, G., P. Ir, C. R. Men, O. O?Donnell, and E. Van Doorslaer (2013). Financial protection of patients through compensation of providers: the impact of health equity funds in cambodia. *Journal of Health Economics 32*(6), 1180–1193.

Foster, A. D. and M. R. Rosenzweig (1995). Learning by doing and learning from others: Human capital and technical change in agriculture. *Journal of Political Economy 103*(6), 1176–1209.

Fraker, T. and R. Moffitt (1988). The effect of food stamps on labor supply: A bivariate selection model. *Journal of Public Economics 35*(1), 25–56.

Fu, H., J. E. Darroch, S. K. Henshaw, and E. Kolb (1998). Measuring the extent of abortion underreporting in the 1995 national survey of family growth. *Family planning perspectives*, 128–138.

Gaarder, M. M., A. Glassman, and J. E. Todd (2010). Conditional cash transfers and health: unpacking the causal chain. *Journal of development effectiveness 2*(1), 6–50.

Gaskell, G. D., D. B. Wright, and C. A. O'Muircheartaigh (2000). Telescoping of landmark events: Implications for survey research. *The Public Opinion Quarterly 64*(1), 77–89.

Gelbach, J. B., J. Klick, and T. Stratmann (2009). Cheap donuts and expensive broccoli: the effect of relative prices on obesity. *Available at SSRN 976484*.

Geroski, P. A. (2000). Models of technology diffusion. *Research policy 29*(4), 603–625.

Gertler, P. (2004). Do conditional cash transfers improve child health? evidence from progresa's control randomized experiment. *The American Economic Review 94*(2), 336–341.

Gertler, P. J., P. I. Giovagnoli, and S. Martinez (2014). Rewarding provider performance to enable a healthy start to life: evidence from argentina's plan nacer. *World Bank Policy Research Working Paper* (6884).

Gertler, P. J. and C. Vermeersch (2012). Using performance incentives to improve health outcomes. *World Bank Policy Research Working Paper* (6100).

Gibbons, R. S. (1997). An introduction to applicable game theory. *Journal of Economic Perspectives 11*(1), 127–149.

Gittleman, M. (2001). Declining caseloads: What do the dynamics of welfare participation reveal? *Industrial Relations: A Journal of Economy and Society 40*(4), 537–570.

Glassman, A., D. Duran, L. Fleisher, D. Singer, R. Sturke, G. Angeles, J. Charles, B. Emrey, J. Gleason, W. Mwebsa, et al. (2013). Impact of conditional cash transfers on maternal and newborn health. *Journal of health, population, and nutrition 31*(4 Suppl 2), S48.

Gleason, P., P. Schochet, and R. Moffitt (1998). The dynamics of food stamp program participation in the early 1990s. Technical report, Mathematica Policy Research.

Gray, P. G. (1955). The memory factor in social surveys. *Journal of the American Statistical Association 50*(270), 344–363.

Griliches, Z. (1957). Hybrid corn: An exploration in the economics of technological change. *Econometrica 25*(4), 501–522.

Griliches, Z. et al. (1986). Economic data issues. *Handbook of econometrics 3*, 1465–1514.

Grogger, J. (2002). The behavioral effects of welfare time limits. *The American economic review 92*(2), 385–389.

Grol, R. (1990). National standard setting for quality of care in general practice: attitudes of general practitioners and response to a set of standards. *Br J Gen Pract 40*(338), 361–364.

Grol, R. (2001). Successes and failures in the implementation of evidence-based guidelines for clinical practice. *Medical care 39*(8), 11–46.

Groves, R. M. (2004). *Survey errors and survey costs*, Volume 536. John Wiley & Sons.

Groves, R. M. (2006). Nonresponse rates and nonresponse bias in household surveys. *Public Opinion Quarterly 70*(5), 646–675.

Groves, R. M. (2011). Three eras of survey research. *Public Opinion Quarterly 75*(5), 861–871.

Groves, R. M. and M. P. Couper (2012). *Nonresponse in household interview surveys*. John Wiley & Sons.

Groves, R. M., F. J. Fowler, M. P. Couper, J. M. Lepkowski, E. Singer, and R. Tourangeau (2009). Survey Methodology. *Wiley Series in Survey Methods. New York: John Wiley and Sons*.

Groves, R. M., F. J. Fowler Jr, M. P. Couper, J. M. Lepkowski, E. Singer, and R. Tourangeau (2011). *Survey methodology*, Volume 561. John Wiley & Sons.

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

Groves, R. M., S. Presser, and S. Dipko (2004). The role of topic interest in survey participation decisions. *Public Opinion Quarterly 68*(1), 2–31.

Guido W. Imbens, J. D. A. (1994). Identification and estimation of local average treatment effects. *Econometrica 62*(2), 467–475.

Gundersen, C. and V. Oliveira (2001). The food stamp program and food insufficiency. *American Journal of Agricultural Economics 83*(4), 875–887.

Hahn, J., P. Todd, and W. Van der Klaauw (2001). Identification and estimation of treatment effects with a regression-discontinuity design. *Econometrica 69*(1), 201–209.

Hall, B. H. and B. Khan (2003). Adoption of new technology. In D. C. Jones (Ed.), *New Economy Handbook*. Academic Press.

Hannan, M. T. and J. Freeman (1984). Structural inertia and organizational change. *American sociological review 49*(2), 149–164.

Hausman, J. (2001). Mismeasured variables in econometric analysis: problems from the right and problems from the left. *The Journal of Economic Perspectives 15*(4), 57–67.

Hausman, J. A., J. Abrevaya, and F. M. Scott-Morton (1998). Misclassification of the dependent variable in a discrete-response setting. *Journal of Econometrics 87*(2), 239–269.

Hawkes, S. J., G. B. Gomez, and N. Broutet (2013). Early antenatal care: does it make a difference to outcomes of pregnancy associated with syphilis? a systematic review and meta-analysis. *PLoS One 8*(2), e56713.

Heckman, J. J. and S. Mosso (2014). The economics of human development and social mobility. Technical report, National Bureau of Economic Research.

Heffetz, O. and K. Ligett (2014). Privacy and Data-Based Research. *The Journal of Economic Perspectives 28*(2), 75–98.

Hirsch, B. T. and E. J. Schumacher (2004). Match bias in wage gap estimates due to earnings imputation. *Journal of Labor Economics 22*(3), 689–722.

Holmstrom, B. and P. Milgrom (1991). Multitask principal-agent analyses: Incentive contracts, asset ownership, and job design. *Journal of Law, Economics, & Organization 7*, 24–52.

Hoynes, H. W. and D. W. Schanzenbach (2012). Work incentives and the food stamp program. *Journal of Public Economics 96*(1), 151–162.

Hudak, B. B., J. O'Donnell, and N. Mazyrka (1995). Infant sleep position: pediatricians' advice to parents. *Pediatrics 95*(1), 55–58.

Huillery, E. and J. Seban (2014). Pay-for-performance, motivation and final output in the health sector: Experimental evidence from the democratic republic of congo. Technical report, Working Paper, Department of Economics, Sciences Po, Paris.

Hyslop, D. R. and G. W. Imbens (2001). Bias from classical and other forms of measurement error. *Journal of Business & Economic Statistics 19*(4), 475–481.

Imbens, G. and K. Kalyanaraman (2011). Optimal bandwidth choice for the regression discontinuity estimator. *The Review of Economic Studies*, rdr043.

Jayachandran, S. (2009). Air quality and early-life mortality evidence from indonesia?s wildfires. *Journal of Human Resources 44*(4), 916–954.

John, L. K., G. Loewenstein, A. B. Troxel, L. Norton, J. E. Fassbender, and K. G. Volpp (2011). Financial incentives for extended weight loss: a randomized, controlled trial. *Journal of general internal medicine 26*(6), 621–626.

Johnson, T. and M. Fendrich (2005). Modeling sources of self-report bias in a survey of drug use epidemiology. *Annals of epidemiology 15*(5), 381–389.

Kahneman, D. (2012). *Thinking, fast and slow*. Macmillan.

Kalwij, A. (2010). An empirical analysis of the association between neighborhood income and unit non-response in the survey of health, ageing, and retirement in europe. *Review of Income and Wealth 56*(2), 351–365.

Kanuk, L. and C. Berenson (1975). Mail surveys and response rates: A literature review. *Journal of Marketing Research*, 440–453.

Karlan, D., M. McConnell, S. Mullainathan, and J. Zinman (2016). Getting to the top of mind: How reminders increase saving. *Management Science*.

Karlan, D. and J. Zinman (2008). Lying about borrowing. *Journal of the European Economic Association 6*(2-3), 510–521.

Katz, J. N. and G. Katz (2010). Correcting for survey misreports using auxiliary information with an application to estimating turnout. *American Journal of Political Science 54*(3), 815–835.

Kirmani, A. and A. R. Rao (2000). No pain, no gain: A critical review of the literature on signaling unobserved product quality. *Journal of Marketing 64*(2), 66–79.

Kish, L. (1965). Survey sampling.

Kolstad, J. T. (2013). Information and quality when motivation is intrinsic: Evidence from surgeon report cards. *American Economic Review 103*(7), 2875–2910.

Krosnick, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied cognitive psychology 5*(3), 213–236.

Lazear, E. P. (2000). Performance pay and productivity. *American Economic Review 90*(5), 1346–1361.

Lazonick, W. (1979). Industrial relations and technical change: the case of the self-acting mule. *Cambridge Journal of Economics 3*(3), 231–262.

Lee, D. S. and T. Lemieux (2010). Regression discontinuity designs in economics. *Journal of Economic Literature 48*(2), 281–355.

Leonard, K. L. and M. C. Masatu (2010). Professionalism and the know-do gap: exploring intrinsic motivation among health workers in tanzania. *Health economics 19*(12), 1461–77.

Levy, D. and J. Ohls (2010). Evaluation of jamaica's path conditional cash transfer programme. *Journal of Development Effectiveness 2*(4), 421–441.

Lillard, L., J. P. Smith, and F. Welch (1986). What do we really know about wages? the importance of nonreporting and census imputation. *The Journal of Political Economy*, 489–506.

Lim, S. S., L. Dandona, J. A. Hoisington, S. L. James, M. C. Hogan, and E. Gakidou (2010). India's janani suraksha yojana, a conditional cash transfer programme to increase births in health facilities: an impact evaluation. *The Lancet 375*(9730), 2009–2023.

Lindbeck, A., S. Nyberg, and J. W. Weibull (1999). Social norms and economic incentives in the welfare state. *Quarterly Journal of Economics 114*(1), 1–35.

Lopez, A., P. Cacoub, I. C. Macdougall, and L. Peyrin-Biroulet (2015). Iron deficiency anaemia. *The Lancet*.

Lyberg, L. and D. Kasprzyk (1991). Data collection methods and measurement error: an overview. *Measurement errors in surveys*, 235–257.

Main, D. S., S. J. Cohen, and C. C. DiClemente (1995). Measuring physician readiness to change cancer screening: preliminary results. *American Journal of Preventive Medicine 11*(1), 54–58.

Mansfield, E. (1961). Technical change and the rate of imitation. *Econometrica 29*(4), 741–766.

Marquis, K. H. and J. C. Moore (1990). Measurement errors in SIPP program reports. Technical report, U.S. Census Bureau.

Martorell, R., P. Melgar, J. A. Maluccio, A. D. Stein, and J. A. Rivera (2010). The nutrition intervention improved adult human capital and economic productivity. *The Journal of nutrition 140*(2), 411–414.

Massey, D. S. and R. 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), 222–236.

Mathiowetz, N., C. Brown, and J. Bound (2001). Measurement error in surveys of the low-income population. In *Studies of welfare populations: Data collection and research issues*, pp. 157–194.

McCrary, J. (2008). Manipulation of the running variable in the regression discontinuity design: A density test. *Journal of Econometrics 142*(2), 698?714.

Meyer, B., R. Goerge, and N. Mittag (2015). *Errors in survey reporting and imputation and their effects on estimates of Food Stamp Program participation*. Working Papers, Harris School of Public Policy.

Meyer, B. D. and N. Mittag (2015). Using linked survey and administrative data to better measure income: Implications for poverty, program effectiveness and holes in the safety net. Technical report, National Bureau of Economic Research.

Meyer, B. D., W. K. Mok, and J. X. Sullivan (2015). Household surveys in crisis. *Journal of Economic Perspectives 29*(4), 199–226.

Miller, G. and K. S. Babiarz (2013). Using performance incentives to improve medical care productivity and health outcomes. Technical report, National Bureau of Economic Research.

Mills, A. (2014). Health care systems in low-and middle-income countries. *New England Journal of Medicine 370*(6), 552–557.

MINSAL (2009a). Informe de gestió plan nacer. Technical report, rea Tcnica, Unidad Ejecutora Central. MINSAL Argentina.

MINSAL (2009b). Informe de gestió plan nacer. Technical report, rea Tcnica, Unidad Ejecutora Central. MINSAL Argentina.

Moffitt, R. A. (2016). Economics of means-tested transfer programs in the united states, volume i.

Mohanan, M., M. Vera-Hernndez, V. Das, S. Giardili, J. D. Goldhaber-Fiebert, T. L. Rabin, S. S. Raj, J. I. Schwartz, and A. Seth (2015). The know-do gap in quality of health care for childhood diarrhea and pneumonia in rural india. *JAMA Pediatrics 169*(4), 349–357.

Moore, J. C. (1998). Self/Proxy Response Status and Survey Response Quality, A Review of the Literature. *Journal of Official Statistics 4*(2), 155–172.

Moss, W., G. L. Darmstadt, D. R. Marsh, R. E. Black, and M. Santosham (2002). Research priorities for the reduction of perinatal and neonatal morbidity and mortality in developing country communities. *Journal of Perinatology 22*(6).

Newman, J. C., D. C. Des Jarlais, C. F. Turner, J. Gribble, P. Cooley, and D. Paone (2002). The differential effects of face-to-face and computer interview modes. *American Journal of Public Health 92*(2), 294–297.

Olson, K. (2006). Survey participation, nonresponse bias, measurement error bias, and total bias. *Public Opinion Quarterly 70*(5), 737–758.

O'Muircheartaigh, C. and P. Campanelli (1998, January). The Relative Impact of Interviewer Effects and Sample Design Effects on Survey Precision. *Journal of the Royal Statistical Society. Series A (Statistics in Society) 161*(1), 63–77.

Organization, W. H. (2006). Standards for maternal and neonatal care: Provision of effective antenatal care. Technical report, World Health Organization, Geneva.

Organization, W. H. (2013). World health statistics. Technical report, World Health Organization, Geneva.

Parente, S. L. and E. C. Prescott (1994). Barriers to technology adoption and development. *Journal of Political Economy 102*, 298–321.

Pathman, D. E., T. R. Konrad, G. L. Freed, V. A. Freeman, and G. G. Koch (1996). The awareness-to-adherence model of the steps to clinical guideline compliance. the case of pediatric vaccine recommendations. *Medical care 34*(9), 873–89.

Phavichitr, N. and A. Catto-Smith (2003). Acute gastroenteritis in children : what role for antibacterials? *Paediatric Drugs 5*(5), 279–290.

Phelps, C. E. (2000). Information diffusion and best practice adoption. *Handbook of Health Economics 1, Part A*, 223 – 264.

Pickery, J., G. Loosveldt, and A. Carton (2001). The effects of interviewer and respondent characteristics on response behavior in panel surveys a multilevel approach. *Sociological Methods & Research 29*(4), 509–523.

Randive, B., V. Diwan, and A. De Costa (2013). India?s conditional cash transfer programme (the jsy) to promote institutional birth: Is there an association between institutional birth proportion and maternal mortality? *PLoS One 8*(6), e67452.

Rasella, D., R. Aquino, C. A. Santos, R. Paes-Sousa, and M. L. Barreto (2013). Effect of a conditional cash transfer programme on childhood mortality: a nationwide analysis of brazilian municipalities. *The lancet 382*(9886), 57–64.

Ridder, G. and R. Moffitt (2007). The econometrics of data combination. *Handbook of econometrics 6*, 5469–5547.

Rosenberg, N. (1972). Factors affecting the diffusion of technology. *Explorations in Economic History 10*, 3–33.

Rosenberg, N. (1982). *Inside the Black Box: Technology and Economics*. Cambridge University Press.

Royer, H., M. F. Stehr, and J. R. Sydnor (2012). Incentives, commitments and habit formation in exercise: Evidence from a field experiment with workers at a fortune-500 company. Technical report, National Bureau of Economic Research.

Rubin, D. B. (1976). Inference and missing data. *Biometrika 63*(3), 581–592.

Ryan, B. and N. C. Gross (1943). The diffusion of hybrid seed corn in two iowa communities. *Rural Sociology 8*, 1–15.

Salm, M. and D. Schunk (2012). The relationship between child health, developmental gaps, and parental education: Evidence from administrative data. *Journal of the European Economic Association 10*(6), 1425–1449.

Schuster, M. A., E. A. McGlynn, and R. H. Brook (1998). How good is the quality of health care in the united states? *The Milbank quarterly 76*(4), 517–563.

Schwarcz, R., A. Uranga, C. Lomuto, I. Martnez, D. Galimberti, O. M. Garca, M. E. Etcheverry, and M. Queiruga (2001). El cuidado prenatal: Gua para la prctica del cuidado preconcepcional y del control prenatal. Technical report, National Ministry of Health, Argentina.

Sharp, J. S. and L. Adua (2010). Examining survey participation and response quality: The significance of topic salience and incentives. *Survey methodology 36*(1), 95–109.

Sirken, M. G. (1999). *Cognition and survey research*, Volume 322. Wiley-Interscience.

Skinner, J. and D. Staiger (2015). Technology diffusion and productivity growth in health care. *Review of Economics and Statistics 97*(5), 951–964.

Solon, G., S. J. Haider, and J. Wooldridge (2013). What are we weighting for? Technical report, National Bureau of Economic Research.

Stampini, M. and L. Tornarolli (2012). The growth of conditional cash transfers in latin america and the caribbean: did they go too far? Technical report, IZA Policy Paper.

Strauss, R. S. and W. H. Dietz (1998). Growth and development of term children born with low birth weight: effects of genetic and environmental factors. *The Journal of pediatrics 133*(1), 67–72.

Sudman, S. and N. Bradburn (1974). *Response effects in surveys. A review and synthesis.* Aldine Publishing Company.

Sudman, S. and N. M. Bradburn (1973). Effects of time and memory factors on response in surveys. *Journal of the American Statistical Association 68*(344), 805–815.

Tamborini, C. R. and C. Kim (2013). Are proxy interviews associated with biased earnings reports? marital status and gender effects of proxy. *Social science research 42*(2), 499–512.

Tamura, T., R. L. Goldenberg, J. Hou, K. E. Johnston, S. P. Cliver, S. L. Ramey, and K. G. Nelson (2002). Cord serum ferritin concentrations and mental and psychomotor development of children at five years of age. *The Journal of pediatrics 140*(2), 165–170.

Taylor, S. E. and S. C. Thompson (1982). Stalking the elusive 'vividness' effect. *Psychological Review 89*(2), 155–181.

Thaddeus, S. and D. Maine (1994). Too far to walk: maternal mortality in context. *Social science & medicine 38*(8), 1091–1110.

Thaler, R. and C. Sunstein (2009). *Nudge: Improving Decisions about Health, Wealth, and Happiness*. Penguin Books, New York.

Thapar, N. and I. R. Sanderson (2004). Diarrhoea in children: an interface between developing and developed countries. *The Lancet 363*(9409), 641–653.

Todorov, A. and C. Kirchner (2000). Bias in proxies' reports of disability: data from the national health interview survey on disability. *American Journal of Public Health 90*(8), 1248.

Tourangeau, R. (1984). Cognitive sciences and survey methods. *Cognitive aspects of survey methodology: Building a bridge between disciplines*, 73–100.

Tourangeau, R., R. M. Groves, and C. D. Redline (2010). Sensitive topics and reluctant respondents demonstrating a link between nonresponse bias and measurement error. *Public Opinion Quarterly*, nfq004.

Tourangeau, R. and K. A. Rasinski (1988). Cognitive processes underlying context effects in attitude measurement. *Psychological bulletin 103*(3), 299.

Tourangeau, R., L. J. Rips, and K. Rasinski (2000). *The psychology of survey response*. Cambridge University Press.

Tourangeau, R. and T. W. Smith (1996). Asking sensitive questions the impact of data collection mode, question format, and question context. *Public opinion quarterly 60*(2), 275–304.

Tourangeau, R. and T. Yan (2007). Sensitive questions in surveys. *Psychological bulletin 133*(5), 859.

Turner, C. F., L. Ku, S. M. Rogers, L. D. Lindberg, J. H. Pleck, and F. L. Sonenstein (1998). Adolescent sexual behavior, drug use, and violence: increased reporting with computer survey technology. *Science 280*(5365), 867–873.

US Census Bureau (2006). *Design and Methodology: Current Population Survey*. US Census Bureau.

US Census Bureau (2008). *Survey of Income and Program Participation: User?s Guide*. US Census Bureau.

US Census Bureau (2014). *American Community Survey: Design and Methodology*. US Census Bureau.

Volpp, K. G., L. K. John, A. B. Troxel, L. Norton, J. Fassbender, and G. Loewenstein (2008). Financial incentive-based approaches for weight loss: a randomized trial. *JAMA 300*(22), 2631–2637.

Volpp, K. G., A. B. Troxel, M. V. Pauly, H. A. Glick, A. Puig, D. A. Asch, R. Galvin, J. Zhu, F. Wan, J. DeGuzman, et al. (2009). A randomized, controlled trial of financial incentives for smoking cessation. *New England Journal of Medicine 360*(7), 699–709.

Wagner, D. and M. Layne (2014). The person identification validation system (pvs): Applying the center for administrative records research and applications?(carra) record linkage software. *Center for Administrative Records Research and Applications Working Paper 1*.

Wooldridge, J. M. (2007). Inverse probability weighted estimation for general missing data problems. *Journal of Econometrics 141*, 1281–1301.

Wooldridge, J. M. (2010). *Econometric analysis of cross section and panel data*. MIT press.