Nudging to Use: Achieving Safe Water Behaviors in Kenya and Bangladesh

Jill Luoto[1], David Levine[2], Jeff Albert[3], and Stephen Luby[4]

Abstract

Convincing people to adopt preventive health behaviors consistently is difficult, yet many lives could be saved if we understood better how to do so. For example, low-cost point-of-use (POU) technologies such as chlorine and filters can substantially reduce diarrheal disease, responsible for nearly 1.7 million child deaths each year. Nonetheless, these products are not consistently used anywhere in the developing world, even when available and heavily subsidized. We ran complementary randomized field studies in rural western Kenya and urban Dhaka, Bangladesh in which households received free trials of POU products to test the role of marketing on usage of these preventive health goods. Health-oriented marketing messages inspired by behavioral economics incrementally increase use of all products in both countries. We discuss how our findings from these two studies complement and contradict each other, and what we can learn generally about the uptake of these and other preventive health goods.

---

* We have benefitted from conversations with Stefano DellaVigna, Ryan Kellogg, Steve Luby, Jeremy Magruder, Ted Miguel, Shanthi Nataraj, Clair Null, Elisabeth Sadoulet, Dean Spears, Alix Zwane, and seminar participants at UC Berkeley, USF and NEUDC. For the Kenya project, we owe many thanks to Davis Ochieng, Salome Aoko, Emeldah Orowe, George Wambiya, Maureen Mbolo, and Susan Leah Osanjo for excellent field data collection, as well as to Brad Lang, Sam Ombeki, Gordon Oluoch, and the CARE-Kenya Kisumu staff for management assistance. In Bangladesh, we are grateful to Mohammad Abdul Kadir, the field team Tahmina Parvin, Fatema Tuj-johra, Rita Begum, Halima Hawa, Kathika Rani Biswas, Abdul Karim, Shahnaj Aktar, supervised by Farzana Yeasmin. We gratefully acknowledge funding from the following sources: Blum Center for Developing Economies and Institute for Research on Labor and Employment at UCBerkeley, the Swedish International Development Agency (SIDA), and the P&G Fund of the Greater Cincinnati Foundation. These studies were partially funded by the P&G Fund of the Greater Cincinnati Foundation, which is associated with the Procter & Gamble Company (the manufacturer of PUR). All errors are our own.

[1]RAND, Santa Monica, CA, email: jluoto@rand.org. [2]Haas School of Business, University of California, Berkeley. [3]Aquaya Institute, San Francisco, CA, [4]Center for Innovation in Global Health, Standford University.

Highlights

- We run complementary field experiments in rural Kenya and urban Dhaka, Bangladesh.
- We provide free safe water products to households in both settings.
- Marketing messages inspired by behavioral economics increase product usage.
- Complementary findings across settings help address concerns of external validity.

I. Introduction

Inadequate access to safe water is a primary cause of the estimated 1.7 million child deaths from diarrhea that occur each year in poor countries (Boschi-Pinto, Velebit, and Shibuya 2008). Fortunately, low-cost point-of-use (POU) safe water technologies such as chlorine or a filter can substantially reduce diarrheal incidence (Clasen et al. 2006; Clasen et al. 2007; Clasen et al. 2005; Clasen et al. 2004; Clasen, Brown, and Collins 2006; Brown, Sobsey, and Loomis 2008; Du Preez et al. 2008; Arnold and Colford, 2007). A recent meta-analysis of 31 randomized trials of POU products yields a pooled estimate of 42% reduction in diarrheal disease risk (Waddington, Snilstveit, and White, 2009). Unfortunately, POU technologies deliver no health benefits unless used, and adoption and regular use of POU technologies remains low among the global poor (Rosa and Clasen, 2009; Kremer et al. 2009; Luoto et al. 2011; Luby et al. 2008). Usage typically remains low even after years of social marketing in some settings (Holla and Kremer, 2009; Kremer et al. 2009).

Economists typically assume that a low adoption rate implies the perceived costs of purchasing the product and effort of using it are greater than the perceived benefits. To increase adoption and usage rates, one then must either lower the costs or increase the perceived benefits. In this case, interventions could consist of free product provision (to lower the costs of purchase; e.g., Dupas 2010; Cohen and Dupas 2010), informational interventions (to increase the perceived benefits of the technology; e.g., Jalan and Somanathan, 2008), and, perhaps, improving the technology itself (to lower the costs or increase the benefits).

This paper presents results from two complementary field experiments conducted in rural western Kenya and the urban slums of Dhaka, Bangladesh that aimed to ease all of these constraints. In both settings, participating households received free trials with a variety of POU products as well as repeated educational messages about the importance of safe drinking water and its link with diarrheal illness.[1] A primary goal of these studies was to determine if alleviating all of these traditional constraints would help spur market development for safe water products.

In both settings, some products were more popular than others and information also played a role, in that households who were informed about contamination in their nearby water source used POU products more than others (see Albert et al. 2010 and Luoto et al. 2011). Nevertheless, even when the price was zero, even among those consumers who correctly identified the causes of diarrhea and the reasons for water treatment, and even among consumers who said they intended to use the product regularly, in neither setting did any product approach universal adoption.

Low usage suggests that behavioral constraints may also affect rates of treating drinking water each day. We hypothesized that decision-making heuristics emphasized by psychologists and behavioral economists as important in other contexts (e.g., Cialdini 1993; Bertrand et al.

[1] A companion paper (Jeff Albert et al. 2010) compares the popularity and usage of the several POU products. Because all the POU products markedly improve the safety of drinking water, we focus here on the non-price barriers that limit their usage. The appropriate role of charging for POU products has been explored by others (e.g., Kremer et al. (2009); Ashraf et al. (2007); Holla & Kremer (2009)).

2009; Thaler and Sunstein 2009; Ariely 2009) can apply to safe water behaviors as well. Specifically, we hypothesized that the framing of the decision to use a POU product might matter and that requesting a verbal commitment to use a POU product could increase regular usage.

We tested the importance of framing by providing different participants with different messages. Some messages emphasized that POU products improved health while others emphasized both avoiding disease and improving health. This intervention was designed to test competing recommendations about effective framing in the psychological and behavior change literatures (PSI 2007; Rothman et al. 1999; Tversky and Kahneman 1981).

We tested the importance of public commitment by asking respondents (in the presence of the enumerator) if they would use the product daily. Again, this intervention builds on an extensive experimental literature in other settings (Webb and Sheeran 2006; Cialdini 1993; Greenwald et al. 1987).

Our replication of a field experiment in two very different settings – one in urban Asia, the other in rural Africa – is (so far as we know) unique for economists. This replication allows us to examine the potential generalizability of our results. The external validity of field experiments and the proper role and ability of RCTs in informing policy has recently come under debate among development economists (Banerjee and Duflo 2008, 2011; Ravallion 2012; Rodrik 2008; Rosenzweig 2012). Replication of influential economic studies is now being actively encouraged,[2] as is a more careful consideration of the external validity of findings from field experiments (Allcott and Mullainathan 2012). Combining our results from two separate RCTs from two very different settings should add weight to our findings, particularly due to the perceived "soft" nature of marketing interventions which are often viewed as being subject to the influences of local context.

In Kenya we observed usage from the freely provided products in 48% of follow-up visits, which was far higher than the 10% rate in Bangladesh. (In both nations rates were lower for having zero detectable *E. coli* and in Kenya self-reported usage was higher, at 67%.) We cannot say with certainty why so many fewer households used their products in Dhaka, but one possibility is that POU products are comparatively less well known in Dhaka where there has been significantly less social marketing of them.

While take-up in the baseline condition differed across nations, in both settings both marketing appeals increase POU usage. Furthermore, the effects of the messages appear additive in both settings, and together they raise rates of observed water treatment by 3-11 percentage points, or 12-50%. Specifically, in Kenya messages that framed safe water technologies as both avoiding disease and improving health (as opposed to the positive frame that mentioned only improving health) raised usage rates by 3-5 percentage points, or 5-13%. The effects of marketing messages that asked consumers to publicly commit to water treatment increased usage

[2] The International Initiative for Impact Evaluation (3ie) has recently announced a grant window inviting teams to replicate existing economic impact evaluation studies whose results are counterintuitive or widely cited. See http://www.3ieimpact.org/en/funding/replication-window/ (accessed 17 July 2012).

by almost exactly the same amount. Results from Bangladesh were broadly similar. If these results generalize, a conservative estimate that only incorporates effects on mortality (and ignores morbidity impacts) suggests that adding these messages to a standard intervention that distributes POU devices for free in settings with high diarrhea-specific child mortality could save approximately six additional disability-adjusted life years (DALYs) per 1000 households per year, with minimal increase in cost.

Our results suggest promising avenues to explore for improvements in the marketing of health products distributed by the private sector. Many of these interventions are potentially cost-effective and necessitate only a rethinking of existing marketing strategies.

2. Study Settings

2.1 Background and Summary Statistics

We partnered with CARE-Kenya from July 2008 to February 2009 to study 400 randomly selected compounds (a collection of households; Luo tradition allows for polygamous marriages) in 28 villages within the largely rural Nyawita sublocation of Nyanza province in western Kenya. Nyanza is among Kenya's poorest regions and was chosen due to its seasonal reliance on turbid earthpans - surface reservoirs that sometimes go dry between rainy seasons - for drinking water. Drinking water conditions vary considerably throughout the year, but most respondents prefer rainwater collection and public taps when available.[3] Other water sources include the Yala River bordering one side of Nyawita and the earthpans that dot the landscape.

In Dhaka, we partnered with ICDDR,B from January – December 2009 to conduct a study among 800 households residing in low-income sections of the densely-populated mixed-income community of Mirpur. This is a crowded urban ("slum") community where the most common source for drinking water at baseline was piped water (74%). Rates of water contamination were comparable to Kenya, likely due to inadequate sanitation and environmental seepage. We selected several neighborhoods within Mirpur that survey staff knew to be relatively poor and whose residents frequently present themselves at ICDDR,B's clinic for diarrheal treatment.

Table 1 presents baseline summary statistics of households from both settings. Both are poor. A similar share of respondents (18%) report an education level beyond primary and average household size is about six people in both countries. More households in Dhaka report an iron roof (92% versus 63%), but more respondents also self-report being illiterate (43% versus 11%). Twice as many Dhaka households had objectively verified soap in the household during the baseline visit (92% versus 45%).

Average available water quality is comparably poor in both settings.[4] In Kenya, 86.5% of household stored water samples taken during the baseline survey (from rain water, tap water,

[3] Nyawita has rainy seasons near April and August (each with a monthly average near 160 mm), with moderate rain in the short dry season (averaging about 100 mm in June) and little rain in the long dry season (averaging less than 40 mm in January)(Kenya Agricultural Research Institute (KARI)).

[4] We can reject the null hypotheses of equal medians and equal distributions of *E. coli* contamination in stored water at the 1% level as based on Wilcoxon and Kolmogorov-Smirnov tests, respectively, when we

earthpan water, and river water) tested positive for *E. coli*. The mean and median *E. coli* counts at baseline were 155 and 19, respectively. As these baseline measures were taken toward the end of the longer rainy season, when rainwater was plentiful, it is likely that, for most of the year untreated water quality is even lower in Kenya. Similarly, in Dhaka 83% of household water samples taken from households without free POU products tested positive for *E. coli* and the mean and median rates of contamination were 182 and 43.5, respectively.

Rates of reported diarrheal prevalence were also high at baseline in both settings, although more prevalent in Kenya; 42% of Kenyan and 17% of Bangladeshi respondents reported that a child under five years old had suffered an episode of diarrhea in the preceding two weeks. Such high baseline prevalence was matched by high rates of reported concern; a majority (55%) of Kenyan respondents freely named diarrhea in their list of the three most problematic diseases affecting their district. This question was slightly rephrased for the Dhaka sample where half (50%) freely named diarrhea as the single most problematic disease affecting their district.

Despite high rates of water contamination, self-reported diarrheal incidence, and expressed concern, in both settings very few households took preventive action, even if they knew of means to do so. Just 18% of Kenyan respondents reported consistently boiling their drinking water even though 58% named "boil drinking water" as a method of diarrhea prevention. 38% of Dhaka respondents named boiling as a means of prevention, and 34% reported boiling their water at baseline.

Furthermore, in Kenya 98% of respondents had heard of at least one point-of-use water treatment method (WaterGuard, a dilute chlorine solution), but only 7% reported that their current drinking water was treated by a POU method at baseline and we could detect chlorine in only 1.5% of homes. In Dhaka there was significantly less baseline awareness of POU products (likely due to their comparatively recent market introduction accompanied by less social marketing), and no household reported usage of any POU method at baseline.

2.2 Kenya's Experimental Design

We describe here the experimental design in Kenya, highlighting relevant differences for the Dhaka study. Details on Dhaka's experimental design can be found in Luoto et al. (2011).

Prior to the start of the Kenya study, CARE staff conducted a census of all compounds in the 28 villages and recorded which had a child under five, the sole criterion for inclusion in the study. From this list, 400 compounds were chosen by a random-number generator. (In Dhaka the baseline sample was 800 households.)

In July-August 2008, our enumerators visited these compounds and asked to conduct a baseline interview with the mother of the youngest child. If that mother was not available, a mother of a child under five was selected. If no eligible mother was available, enumerators were allowed to substitute the father (11% of baseline interviews). In the rare instances that no one in

---

compare baseline levels of stored water quality in Kenya with stored water quality from samples taken from households assigned to the control group in Dhaka (i.e., water quality in the absence of free POU products for both settings).

the household was available or willing to participate in the study, we substituted the nearest compound with a child under five on the basis of logistical convenience due to the dispersed rural landscape.

The baseline interview asked respondents about their current water and hygiene knowledge and behaviors, as well as prior exposure to any POU technologies. Enumerators then read an educational script about the dangers of unsafe drinking water, followed by detailed presentations on three POU products in randomized order: a liquid chlorine product branded as WaterGuard, Procter & Gamble's flocculant-disinfectant powder branded as PUR, and a gravity-driven porous ceramic filter. All three products have been shown to substantially reduce contamination in drinking water in a number of previous randomized trials (Clasen, et al. 2006; Clasen et al. 2005; Jain et al. 2010; Crump et al. 2004).[5] (In Dhaka we added a tablet chlorine product and replaced the ceramic filter with a siphon filter; see webappendix.)

Enumerators then presented marketing messages with randomly assigned frames. Half of the households heard a "positively framed" message that emphasized the gains from POU usage while the other half were given a "contrast frame" that contrasted what one stands to lose from non-use with the gains from POU product usage.

At the end of the baseline interviews, respondents were randomly assigned one of the three POU technologies for a two-month trial. Each family received a covered bucket with tap along with their assigned product to minimize recontamination of treated water within the household. (In Dhaka, 200 of the 800 households were assigned to a control group and did not receive any products. Also in Dhaka, households typically prefer to use a traditional storage device with a narrow mouth, so we did not distribute safe storage containers to the 600 households in the intervention group that received products.)

At this point, enumerators asked one-half of respondents to commit verbally to the enumerator to use their assigned POU product for all of their drinking water (this was not done for the 200 control households in Dhaka). The framing and commitment treatments were assigned orthogonal to each other and to the assigned product.

Two months later, enumerators revisited all households to collect stored water (both treated and untreated, if available) to test for product usage. Enumerators then re-delivered the same marketing treatment or treatments; that is, positive or contrast framing and in half the case also the commitment script. Enumerators then distributed one of the remaining POU products for a new two-month trial and collected any leftover supplies of the previously assigned product. (The 200 control households in Dhaka were visited every two months solely to collect water samples.) This process was repeated until every participant in the treatment group had experienced a two-month trial of each of the three POU products in random order (four POU products in Dhaka). A timeline of data collection activities for both studies is shown in Figure 1.

[5] See the web appendix for brief introductions to WaterGuard, PUR, and the filter, as well as the tablet chlorine product Aquatabs that was included only in the Bangladesh study. The companion papers, Albert et al. 2010; Luoto, et al. 2011; and Luoto et al. 2012 examine usage rates by product in greater detail for both experiments.

Our studies' two-month product trials before measuring usage is longer than the exposure in many epidemiological studies on POU products (e.g., Crump et al., 2004 and Quick et al., 2002 visited weekly). We believed that the 2-month period would be enough time for households to fall into a stable pattern of usage in order to gauge the ability of our marketing interventions to affect adherence.

3. Marketing and Information Interventions

As in Bertrand et al. (2009) and Dupas (2009), we drew from the extensive decision-making literatures from behavioral economics and psychology to choose interventions that we felt were likely to affect POU usage. While the cited studies examine the decision to purchase a good or service, we are closer to Kremer et al. (2009) in testing how social marketing can affect continued usage of a preventive good. Both of our marketing treatments emphasized the expected health benefits from use of any safe water product, and did not present relative comparisons of one product versus another.

3.1 Framing

The literature offers competing hypotheses on whether framing POU adoption as a gain or a loss should bring about the larger response.

A subset of psychology research (e.g., Rothman et al. (1999)) suggests that a positive frame works best to change behavior. This view has become conventional wisdom among many social marketers. For example, PSI, the world's leader in social marketing, markets and distributes both WaterGuard and PUR in Kenya and 20 other countries. Their manual on social-marketing best practices argues for positively framing messages:

> Consumers need to be inspired by the images and messages they see and hear and then aspire to create the same images in their homes. To create the aspiration, branded campaigns need to focus on the positive attributes of using the safe water solution. To get across the notion that the product can help protect children's health, campaigns must convey images of happy, healthy families that successfully use the product (PSI 2007).

However, due to loss aversion, gains relative to the status quo are often valued less than avoiding losses relative to the status quo (Kahneman and Tversky 1979; Tversky and Kahneman 1981). Moreover, some studies of behavior change argue for use of a "contrast" frame that first describes a problem (e.g., unsafe water) as a loss and then presents a solution (e.g., POU products), emphasizing the individual's ability to achieve the solution (Gass and Seiter 2007).

To test the role of framing, at each survey round enumerators read messages with either a positive or a contrast frame. Respondents in households assigned to the positive frame saw images of smiling children and a visibly clean glass of water as the enumerator read a script about what users stood to gain from regular use of a safe water product. The other respondents saw photographs of a crying child and a visibly dirty glass of water next to a smiling child with a visibly clean glass of water. The accompanying script began by emphasizing that the sad child

had diarrhea from drinking contaminated water. It then became exactly the same as the positively framed message.[6]

3.2 Consistency with Public Commitment

In psychology, the "commitment consistency" theory posits that people will go to great lengths to stay true to a commitment they have made in order to be—or appear to be—consistent (Greenwald et al. 1987; Cialdini 1993). This effect is strongest for commitments made in front of others (Cialdini 1993). There is also evidence that predicting one's own future behavior can influence that behavior (Cialdini 1993; Webb and Sheeran 2006).

However, the strength of this commitment effect outside the laboratory remains inconclusive and may depend on the context.[7] In a meta-analysis of 47 randomized trials, Webb and Sheeran (2006) found verbal commitments consistently increase participants' intentions to act, but (because not all intentions lead to action) have only small to medium effects on actions.

Some people who intend to act then forget to do so (Gollwitzer & Sheeran 2006). Thus, our commitment arm also included a reminder poster. Due to limited sample size, we were unable to test the effects of the reminder poster separately from the effects of the public commitment, so our estimated "commitment" effect also includes any increase from the reminder poster.

Enumerators implemented this treatment for a randomly chosen half of participant households in both countries. The enumerator first asked the respondent if she or he intended to use the assigned POU technology. The enumerator then asked the respondent to promise aloud to use the safe water product to keep their families healthy. This pledge was optional, but all respondents were willing to make it. The respondent was next asked to predict if she or he would be found to be using the safe water product two months later when the enumerator returned. At this point, the respondent was given a poster to hang in their homes as a reminder of the commitment. At the baseline visit, these photographic reminders were posters showing images of all three of the safe water products as well as images of smiling mothers and children. After the first two-month trial with a product, enumerators gave respondents a personalized poster showing images of the products as well as a photo of the respondent herself, taken by the enumerator at baseline. [8]

---

[6] Translations of the verbal scripts and accompanying images for both frames can be found in the online Appendix. These were identical in Dhaka except the images included people of South Asian versus African descent. We gratefully acknowledge input on parts of the positively framed verbal script from members of the Rural Water Project (RWP) in Busia, Kenya, and ideas from Meyerowitz & Chaiken (1987) and Block & Keller (1995).

[7] For example, Greenwald et al. (1987) find that a spoken commitment to vote has a positive effect on voting, but Smith et al. (2003) do not. In poor settings, Kremer & Miguel (2007) find no effect of adolescent respondents in Kenya committing to taking a deworming drug and Dupas (2009) finds no effect of a verbal commitment on subsequent purchases of mosquito nets.

[8] Figure A.3 and A.4 in the online Appendix shows the poster delivered to homes in the commitment treatment during the baseline visit and a personalized poster delivered at the first follow-up. We also delivered personalized posters to the "control" households at the final interview because the posters became valued in the communities in both countries. We thank Clair Null for suggesting the poster.

4. Data Description and Summary Statistics

4.1 Water Collection and Measuring Product Usage

We describe the water collection in the Kenya study, and point out key differences for the Dhaka study.

At each (unannounced) household visit, enumerators performed a variety of tests to measure water quality and product usage. They drew samples from a household's stored supply of drinking water, which we tested for fecal contamination, as indicated by the presence of *E. coli*. Our POU products can treat up to 20 liters of drinking water at a time, but Kenyan villagers often collect more than that, so it was common to find both untreated and treated drinking water on hand in a household. In such cases, we drew samples from both. (Dhaka households almost always have a nearby tap, so there is less stored water at home. Thus, we drew a single water sample from the water they designated as for drinking.)

We analyze four measures of product usage. Our first indicator is self-reported product usage. Self-reports are comparable across products, but are likely to overestimate actual usage due to courtesy bias (Kremer et al. 2009; Jain et al. 2010). We also create an indicator for a household having a positive chlorine test for the chlorine-based products, or an indicator for the enumerator observing usage in the case of the filter. This definition of usage is likely to be a lower bound in the case of the chlorine-based products due to the dissipation of chlorine over time (Kremer et al. 2009). Our third definition of usage is an indicator dummy for the household's drinking (defined as a household's treated water if present; otherwise its untreated water) water having no detectable *E. coli*; that is, less than 1 coliform forming unit [CFU] per 100 mL of water, a level which World Health Organization (WHO) guidelines categorize as indicating no risk of contracting illness (WHO 1997). Our final definition of usage is a continuous measure: the natural log of the count (most probable number, or MPN) of *E. coli* (CFU/100 mL) in a household's drinking water. We code $\log_{10}(E.\ coli=0)$ as -1 and, for households with *E. coli* counts above the maximum detectable value (2419.6 CFU/mL), we substituted that maximum value. For this usage measure, smaller (or more negative) values imply greater usage.

Water treated with a POU product can have detectable *E. coli* if the product is not highly effective (usually a result of user error or recontamination). Thus, all of the marketing randomizations were implemented orthogonally to product assignments; product efficacy should not affect these results. Nonetheless, we include product fixed effects in some models to control for differences in rates of product usage and of their effectiveness when used.

Conversely, if the source water is not contaminated and no contamination occurs during household storage, then even untreated water may have no detectable *E. coli*. Thus, we compare measures of contamination in a household's treated drinking water to its pre-treated or source water in Kenya. In Dhaka, we compare water quality from intervention households with samples

taken from the 200 control households that were not provided free POU products in an intention to treat (ITT) analysis (see Figure 1).

4.2 Balance and Attrition

In both nations baseline characteristics do not predict treatment status (see appendix).

Over the eight months of the Kenya study, 30 of the original 400 households dropped out, resulting in an overall retention rate of 92.5%. Between each successive full round of surveys, retention rates were 97%, 98% and 98%, respectively. When we run a probit regression that predicts dropout as a function of all treatment assignments in Kenya, the Chi-squared test p-value is .16 (estimation not shown). By far the most common reason for a household to drop out of the study was migration to an urban area; thus, our Kenya results are most representative of a persistently rural population. We include all household-survey wave observations when we present our main estimation for effects of the marketing interventions (see Figure 1). Results on the subsample of those who do not attrit are identical, and available from the authors upon request.

In Dhaka, we found 575 of the 600 intervention households (755 of the original 800 participant households) at the 8-month exit survey. The most common reason for a household to drop out of the study was outmigration from the community. Attrition does not appear related to a household's first assigned products or other randomized treatment assignments. When we ran a probit regression predicting dropout as a function of all treatment assignments in the Dhaka sample, the joint Chi-squared test was not statistically significant (p-value = 0.24). Again, we include all household-survey wave observations in our estimation of effects from the marketing interventions on POU product usage, but results are identical if we restrict to the 575 households that completed the study (see Figure 1).

5. Estimation Strategy

To measure the effects of our marketing interventions on usage, for each country we combine all two-month follow-up survey waves when households had free products and estimate:

$$1.\ Y_{iqt} = \alpha + \beta_F\, T_i^F + \beta_C\, T_i^C + \varepsilon_{iqt}$$

where Y_{iqt} is a measure of usage of product q at time t by household i, T_i^F is an indicator that equals 1 if a household is assigned to receive the "contrast" frame (both positive and negative), and 0 otherwise (that is, if they received only the positive frame), and T_i^C is an indicator for the commitment treatment. Due to the randomized assignment to treatment for both marketing interventions, both coefficients should deliver unbiased estimates of their average causal effects on product usage across all post-baseline waves and products. We cluster the error terms ε_{iqt} in Equation 1 at the household to allow for autocorrelated outcomes across survey waves for the same household and due to the household-level randomization of the marketing treatments. We estimate Equation 1 using both a linear model and maximum likelihood methods; results are identical so we present the linear results for ease of interpretation.

To perform tests of statistical significance across Kenya and Dhaka we also estimate a joint model that interacts country dummies with the marketing treatments:

$$2.\ Y_{iqte} = \alpha + \gamma K_e + \beta_F T_i^F + \beta_{Fe} (T_i^F \times K_e) + \beta_C T_i^C + \beta_{Ce} (T_i^C \times K_e) + \delta_{qe} + \varepsilon_{iqte}$$

where K_e is a dummy indicating if an observation is from Kenya's experiment (versus Dhaka's), and experiment is denoted by e. We also now include product-country fixed effects δ_{qe}.

6. Results

6.1 Effects of Providing Free POU Products

Figure 2 presents several measures of product usage for both settings, averaging across all marketing treatments, products and survey rounds. (See Albert et al. (2010) and Luoto et al. (2011) for a discussion of differences in rates of usage among products for Kenya and Dhaka, respectively.) Here we highlight overall differences between countries and experiments from a nominally identical intervention – giving away free safe water products along with repeated educational information about the importance of safe drinking water.

The top half of Figure 2 presents results for Kenya where we see all measures of usage show large increases in rates of water treatment relative to baseline. This is a pre-post comparison for Kenya and thus cannot account for differences in seasonal water quality (although if we make a similar comparison of pre-treated water to treated water for the indicator of no detectable *E. coli*, results are very similar; also, baseline water quality is from the rainy season when water quality is generally better, so it is likely the differences shown here may be lower bounds). Nonetheless, the free provision of products clearly increases water treatment. As expected, self-reported usage is highest, with an average of 72% of households self-reporting current use of their POU product two months after receiving it across all three products tested in Kenya.

The bottom half of Figure 2 presents identical results from Dhaka. When we compare water quality in intervention households with freely provided POU products versus control households that were not given any products, we see increases in overall water quality but of much smaller magnitudes than in Kenya. This comparison in Dhaka can be attributed as causal due to the randomized nature of treatment status.

6.2 Marketing Effects

Table 2 contains results from estimation of equation 1 on the impacts of the marketing treatments separately for Kenya (panels A) and Dhaka (panel B). In panel C, we present results of cross-equation tests of significance from estimation of equation 2 on the pooled sample and fully interacted model with product-country fixed effects. We consider the joint test for the sum of interventions (that is, the sum of effects from the contrast frame and commitment treatments in Kenya, and the sum of the contrast frame and commitment effects in Dhaka) to be the main result of this paper. We can reject the null hypothesis that the sum effect of the two marketing interventions (framing and commitment) are jointly zero in both countries for all four definitions of usage at the 10% level or greater. We interpret this as evidence that our marketing

interventions caused more households to adhere to water treatment behaviors two months later during follow-up survey rounds.

In Table 3, we breakdown these results and take advantage of the randomized and orthogonal assignment to the marketing interventions to incrementally consider the effects from each of the marketing treatments individually for each country. That is, we present a basic nonparametric cross-tabulation of mean usage rates for households that have been given free POU products but have received different combinations of marketing messages - either the "traditional marketing" treatment of only a positively framed message (which we treat as our control condition), or just the contrast frame or just the commitment treatment, or the combined effects of receiving both marketing interventions (contrast frame and commitment).

Table 3 shows that each marketing message individually increases product usage by a few percentage points in both countries - we successfully *nudge* with each treatment, and the combined effect of the marketing messages stands out as realizing the largest gains in both countries. Together, the marketing treatments increase product usage in Kenya by 6-11 percentage points, or 12-31%, across all measures of usage at all follow-up surveys relative to the "traditional marketing" condition (these increases are statistically significant for having no detectable *E. coli* in column 1 (1% level based on F-statistic), and self-reporting product usage in column 3 (10% level based on F-statistic)). Despite the much lower base rates of product take-up in Dhaka, the combined effect of both marketing interventions raises rates of usage by similar and statistically significant 3-8 percentage points across the various indicators of usage, or 30-50% beyond that achieved by their free provision alone. We find no evidence of substitution effects between marketing messages in either country (see Table 2, or Table A2 in web appendix where we interact the two marketing treatments).

6.2.1 Contrast Frame

Results in Tables 2 and 3 show that the contrast framed message appears more successful in Kenya than in Dhaka. In Kenya, although the effect is not statistically significant for all definitions of usage, the sign and magnitude suggest that contrasting what one stands to lose from nonuse with what one stands to gain from use is more effective than focusing solely on the potential gains. Moreover, the point estimates from Kenya can be fairly large: contrast-frame households are five percentage points more likely to have uncontaminated treated water at home two months later (p-value of 0.06), a nearly 15% increase over usage in households with only a positive frame (Table 2, panel A, column 1).

The weaker results from the framing treatment in Dhaka are disappointing (Table 2, panel B), but do not contradict our findings in Kenya. It is not surprising that different marketing messages would realize different success in different settings, and points to the importance of testing multiple messages in multiple settings. We discuss these points further in the conclusion.

Although the hypothesis motivating the contrast frame was that loss aversion would spur greater action, the results in Tables 2 and 3 are consistent with other interpretations, such as a model of limited attention in which a reminder of sickness adds salience to the treatment

decision by causing people to consider the full spectrum of possible outcomes (DellaVigna 2009).

6.2.2 Consistency with Public Commitment

Whereas the contrast framing treatment appeared relatively more successful in Kenya at nudging households to use their free POU products, the commitment treatment appears comparatively more successful among the Dhaka households. There, the commitment treatment increases the likelihood of a household having no detectable *E. coli* by five percentage points, about 20% more than in non-commitment homes (Table 2 panel B column 1, significant at 5%). Although we suspected that the estimated treatment effect would be inflated by courtesy bias for rates of self-reported usage, results with this measure do not differ substantially (Table 2 column 2 panels A and B, p-value of .16 in Kenya and significant at 1% in Dhaka).

6.3 Robustness Checks

To adjust for any random differences in baseline characteristics that might affect our findings, we ran multivariate regressions controlling for gender, education level and basic wealth indicators such as presence of an iron roof. Results for both countries were very similar to those presented, with slightly improved precision (Table A1). Results for the marketing treatments in Kenya are also similar if we cluster disturbance terms at the village level to allow for correlation across households who may share a water source.

In Table 4 we also estimated a household-level ordered logit that counts the total number of products a household used out of the three total in Kenya, and four total in Dhaka. Results again are very similar to those in tables 2 and 3.

In Kenya, the contrast frame and commitment treatment improve treated water quality by increasing POU product usage, not by improving the quality of source water collected such as by causing shifts to safer water sources. There is no detectable effect of either marketing intervention on untreated water quality in placebo regressions (Table 5). (A similar test is not possible in Dhaka since only one sample of water was collected at household visits.)

We estimate the effects of our marketing interventions separately by product by interacting product dummies with indicators for the marketing treatments (see webappendix Table A3). We lack power to say much about the product-specific effects of our marketing interventions, but no single product is consistently driving these results.

Finally, we adapt a simple empirical test for the potential external validity of our results as suggested by Allcott and Mullainathan (2012) that examines treatment effect heterogeneity within "sub-sites" of a given sample. Specifically, our Kenya data implemented the commitment and contrast framing marketing treatments within 28 villages. If the treatment effects are similar across the 28 villages, there is no evidence against generalizing to other village in rural western Kenya. Similarly, if results are similar in Kenya and Bangladesh, it is more likely that results generalize to other poor settings.

To test these predictions, we first treat each of the 28 villages as a sub-site and regress self-reporting having chlorinated or filtered water at follow-up survey waves on a series of village

dummies interacted with treatment indicators. We also as controls a series of educational and wealth indicators interacted with the treatment dummies and baseline water quality variables.

An F-test on the joint hypothesis of no treatment effect heterogeneity across the "sub-sites" (the series of interacted village and treatment dummies) has p=0.76 for the framing treatment and p=0.66 for the commitment treatment (results available upon request). If we pool our data across countries and perform a similar test with Kenya and Bangladesh as "sub-sites," we again get p-values of around 0.5. Although these tests are only suggestive, we take them as another encouraging sign that our results may be generalizable.

7. Discussion

Our two sister experiments considered the role of marketing messages informed by behavioral economics and psychology to increase usage of freely provided POU safe water products. We find positive and incremental effects from all of our treatments in two very different settings and across a variety of water treatment products.

Despite such similar effects from our "soft" marketing treatments, we find very different base rates of product adoption in response to the more or less identical treatment of free product provision. We cannot say with certainty why so many fewer households used their products in Dhaka, but there are a variety of possible explanations. First, these products are comparatively newer and less well known in Dhaka (Table 1), where there has been significantly less social marketing of them. Also, more households in Kenya rely on non-piped water sources, particularly at various times during the year when the pipes may run dry, and it is easy to imagine that the necessity for treating one's water is more salient when that water is turbid or collected from the same source as where farm animals are brought to drink (although we lack power to test this hypothesis).

Although our experiments were not designed to test for differential take-up across settings, we take this as evidence in favor of the recent push for more replication of economic experiments and a greater consideration of external validity.

We did not measure usage for periods longer than two months. Nonetheless, we are cautiously optimistic on a few fronts. First, our marketing results could potentially be incorporated into preexisting distribution and marketing models for these products at little to no cost. Consumable POU products such as WaterGuard, Aquatabs and PUR are currently packaged and sold in units that are not meant to last more than two months, and are often sold by community health workers who make home visits in a similar manner to that done by our enumerators. In addition, in both settings the two marketing interventions appear additive, which suggests that harnessing even more behavioral principles might realize larger effects.

If we combine our findings from the marketing treatments with those from previous studies, we can roughly estimate expected health effects from the free provision of these POU products as well as the additional health benefits from greater usage due to the marketing treatments. Under reasonable but conservative assumptions (summarized in the webappendix) and focusing solely on averted mortality and not morbidity outcomes for children aged 1-59 months (ignoring neonatal mortality), free provision of POU products to 1000 Kenyan households in Nyanza

province for one year *without* using our preferred marketing messages would save 0.7 lives of children 1-59 months old or approximately 21 disability-adjusted life years (DALYs). Assuming a cost of $4.10 per household-year (Clasen, 2008 Table 2.5b), free chlorine provision to 1000 households for one year would cost roughly $198 per 1-59 month old DALY averted, far less expensive than the WHO definition of "very cost effective" interventions of averting one DALY for less than the gross domestic product per capita (estimated to be $456 in 2008 for Kenya in constant 2000 US dollars[CMH 2001]). Adding our marketing treatments to the regular free chlorine provision in Kenya would increase those savings by roughly 30%, to 0.9 lives saved and 27 DALYs averted among 1-59 month olds, at a cost of $153.62 per DALY.

In urban Dhaka, a substantially lower share of households use chlorine, and fertility and diarrhea-specific child mortality rates are significantly lower than in rural Kenya. Free provision of chlorine with or without our marketing treatments to 1000 households for one year would not be classified by the WHO as "very cost effective" (see webappendix for details).

These calculations are very rough and the variation in usage across nations emphasizes caution in generalizing. We interpret the different conclusions regarding a policy of free chlorine provision across settings as further evidence of the importance of considering the generalizability of any experiment's findings, but highlight that in certain settings such a policy can be potentially very cost-effective, and even more so if coupled with successful nudges to increase rates of their usage. Moreover, the complementary findings from our "soft" marketing interventions across two very different settings suggests that marketing messages informed by well-known principles from behavioral economics and the psychology of decision-making can, under appropriate circumstances, promote development aims. We are excited about the potential generalizability of these findings to other settings and to other preventive health behaviors.

In this paper we focused on the impacts of marketing treatments on usage of safe water products during free trials. We did not measure actual purchases and the study was not powered to measure health effects. Our findings also highlight the heterogeneity in consumer take-up of health prevention measures. More research is needed on these points, as well as to explore what (if any) combinations of messages and interventions can change behavior reliably and consistently.

References

Albert, Jeff, Jill Luoto, and David I Levine. 2010. "End-User Preferences for and Performance of Competing POU Water Treatment Technologies among the Rural Poor of Kenya."

Environment, Science and Technology 44(12):4426–4432. Retrieved (http://www.ncbi.nlm.nih.gov/pubmed/20446726).

Allcott, Hunt, and Sendhil Mullainathan. 2012. "External Validity and Partner Selection Bias." *NBER Working Paper No. 18373*. Retrieved (http://www.stanford.edu/group/SITE/SITE_2012/2012_segment_5/2012_segment_5_papers/allcott.pdf).

Ariely, D. 2009. *Predictably Irrational: The Hidden Forces that Shape our Decisions*. Revised an. New York: Harper Collins Publishers.

Arnold, Benjamin F, and Jack M Colford. 2007. "Treating water with chlorine at point-of-use to improve water quality and reduce child diarrhea in developing countries: a systematic review and meta-analysis." *American Journal of Tropical Medicine and Hygiene* 76(2):354–364.

Ashraf, Nava, James Berry, and Jesse Shapiro. 2010. "Can Higher Prices Stimulate Product Use? Evidence from a Field Experiment in Zambia." *American Economic Review* 100(5):2383–2413.

Banerjee, A, and E Duflo. 2008. "The Experimental Approach to Development Economics." *working paper*. Retrieved (http://economics.mit.edu/files/3158).

Banerjee, A, and E Duflo. 2011. *Poor Economics: A Radical Rethinking of the Way to Fight Global Poverty*. New York: Public Affairs.

Bertrand, Marianne, Dean Karlan, Sendhil Mullainathan, Erik Shafir, and Jonathan Zinman. 2009. "What's Advertising Content Worth? Evidence from a Consumer Credit Marketing Field Experiment (January 23, 2009)." *Yale University Economic Growth Center Discussion Paper No. 968; Yale Economics Department Working Paper No. 58.*

Block, L G, and P A Keller. 1995. "When to Accentuate the Negative: The Effects of Perceived Efficacy and Message Framing on Intentions to Perform a Health-Related Behavior." *Journal of Marketing Research* 32 (May):192–203.

Boschi-Pinto, Cynthia, L Velebit, and K Shibuya. 2008. "Estimating child mortality due to diarrhoea in developing countries." *Bulletin of the World Health Organization* 86(9):710–717. Retrieved July 26, 2011 (http://www.who.int/bulletin/volumes/86/9/07-050054.pdf).

Brown, J, and M Sobsey. 2006. "Independent Appraisal of Ceramic Water Filtration Interventions in Cambodia: Final Report." *Report submitted to UNICEF*.

Brown, Joe, Mark D Sobsey, and Dana Loomis. 2008. "Local Drinking Water Filters Reduce Diarrheal Disease in Cambodia: A Randomized, Controlled Trial of the Ceramic Water Purifier ." *The American Journal of Tropical Medicine and Hygiene* 79 (3):394–400. Retrieved (http://www.ajtmh.org/content/79/3/394.abstract).

Chiller, T et al. 2006. “Reducing Diarrhoea in Guatemalan Children: Randomized Controlled Trial of Flocculant- Disinfectant for Drinking-Water.” *Bulletin of the World Health Organization* 1.

Cialdini, R B. 1993. *Influence: The Psychology of Persuasion*. New York: Morrow.

Clasen, T, G Garcia Parra, S Boisson, and S Collin. 2005. “Household-based Ceramic Water Filters for the Prevention of Diarrhea: A Randomized, Controlled Trial of a Pilot Program in Colombia.” *American Journal of Tropical Medicine and Hygiene* 3(4):790–795.

Clasen, T, I Roberts, T Rabie, W Schmidt, and S Cairncross. 2006. “Interventions to Improve Water Quality for Preventing Diarrhoea (Review).” *The Cochrane Library* 4.

Clasen, Thomas F, Joseph Brown, and Simon M Collin. 2006. “Preventing diarrhoea with household ceramic water filters: Assessment of a pilot project in Bolivia.” *International Journal of Environmental Health Research* 16(3):231–239. Retrieved (http://www.tandfonline.com/doi/abs/10.1080/09603120600641474).

CLASEN, THOMAS F, JOSEPH BROWN, SIMON COLLIN, OSCAR SUNTURA, and SANDY CAIRNCROSS. 2004. “REDUCING DIARRHEA THROUGH THE USE OF HOUSEHOLD-BASED CERAMIC WATER FILTERS: A RANDOMIZED, CONTROLLED TRIAL IN RURAL BOLIVIA.” *The American Journal of Tropical Medicine and Hygiene* 70(6):651–657. Retrieved (http://www.ajtmh.org/content/70/6/651.abstract).

Clasen, Thomas, Wolf-Peter Schmidt, Tamer Rabie, Ian Roberts, and Sandy Cairncross. 2007. “Interventions to improve water quality for preventing diarrhoea: systematic review and meta-analysis.” *BMJ (Clinical research ed.)* 334(7597):782. Retrieved August 22, 2011 (http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1851994&tool=pmcentrez&rendertype=abstract).

Clasen Tom, Laurence Haller. 2008. “Water Quality Interventions to Prevent Diarrhoea: Cost and Cost-Effectiveness.” *World Health Organization Public Health and the Environment*.

CMH. 2001. *Commission on Macroeconomics and Health, Macroeconomics and health: investing in health for economics development.* Boston: Center for International Development at Harvard University.

Cohen, J, and P Dupas. 2010. “Free Distribution or Cost Sharing? Evidence from a Randomized Malaria Prevention Experiment.” *Quarterly Journal of Economics* CXXV(1).

Crump, J A et al. 2005. “Household Based Treatment of Drinking Water with Flocculant-Disinfectant for Preventing Diarrhea in Areas with Turbid Source Water in Rural Western Kenya: Cluster Randomized Controlled Trial.” *British Medical Journal* 331(7515):478–484.

DellaVigna, Stefano. 2009. “Psychology and Economics: Evidence from the Field.” *Journal of Economic Literature* 47:315–372.

DHS. 2009. *Kenya Demographic and Health Survey 2008-2009 Final Report*.

Dupas, Pascaline. 2009. “What Matters (and What Does Not) in Households’ Decision to Invest in Malaria Prevention?” *American Economic Review: Papers and Proceedings* 99(2):224–230.

Dupas, Pascaline. 2010. “Short-Run Subsidies and Long-Run Adoption of New Health Products: Evidence from a Field Experiment.” *NBER Working Paper No. 16298* (August version). Retrieved (http://www.stanford.edu/~pdupas/Subsidies&Adoption.pdf).

Fischer Walker, Christa L, Martin J Aryee, Cynthia Boschi-Pinto, and Robert E Black. 2012. “Estimating Diarrhea Mortality among Young Children in Low and Middle Income Countries.” *PLoS ONE* 7(1):e29151. Retrieved (http://dx.doi.org/10.1371%2Fjournal.pone.0029151).

Gass, R H, and J Seiter. 2007. *Persuasion, Social Influence, and Compliance Gaining*. Third. Pearson Education, Boston, MA.

Gollwitzer, PM, and P Sheeran. 2006. “Implementation Intentions and Goal Achievement: A Meta-Analysis of Effects and Processes.” *Advances in Experimental Social Psychology* 38:69–119.

Greenwald A. G., C G Carnot, R Beach, and B Young. 1987. “Increasing voting behavior by asking people if they expect to vote.” *Journal of Applied Psychology* 72:315–318.

Holla, Alaka, and Michael Kremer. 2009. “Pricing and Access: Lessons from Randomized Evalutions in Education and Health.” *Center for Global Development working paper* (158).

Hutton, G, and L Haller. 2004. “Evaluation of the Costs and Benefits of Water and Sanitation Improvements at the Global Level.” in *World Health Organization, Geneva*.

Jain, Seema et al. 2010. “Sodium Dichloroisocyanurate Tablets for Routine Treatment of Household Drinking Water in Periurban Ghana: A Randomized Controlled Trial .” *The American Journal of Tropical Medicine and Hygiene* 82 (1):16–22. Retrieved (http://www.ajtmh.org/content/82/1/16.abstract).

Jalan, Jyotsna, and E Somanathan. 2008. “The importance of being informed: Experimental evidence on demand for environmental quality.” *Journal of Development Economics* 87:14–28.

Kahn, James G et al. 2012. “Integrated HIV Testing, Malaria, and Diarrhea Prevention Campaign in Kenya: Modeled Health Impact and Cost-Effectiveness.” *PLoS ONE* 7(2):e31316. Retrieved (http://dx.doi.org/10.1371/journal.pone.0031316).

Kahneman, Daniel, and Amos Tversky. 1979. “Prospect Theory: An Analysis of Decision under Risk.” *Econometrica* 47 (2):263–291.

Kenya Agricultural Research Institute (KARI), Kibos Station. n.d. “Unpublished data.”

Kremer, M, E Miguel, S Mullainathan, C Null, and A Zwane. 2009. “Trickle Down: Chlorine Dispensers and Household Water Treatment.” *working paper*. Retrieved (http://www.economics.harvard.edu/files/faculty/36_WG_2011_CLEAN_04_14_MKWITHOUT COMMENTS.pdf).

Kremer, Michael, and Edward Miguel. 2007. “The Illusion of Sustainability.” *The Quarterly Journal of Economics* 122(3):1007–1065.

Lantagne, D S. 2001. “Investigation of the Potters for Peace Colloidal Silver-Impregnated Ceramic Filter: Intrinsic Effectiveness and Field Performance in Rural Nicaragua.” *Report prepared for USAID, Washington, D.C.*

Luby, S et al. 2001. “A Low-Cost Intervention for Cleaner Drinking Water in Karachi, Pakistan.” *International Journal of Infectious Diseases* 5:144–150.

Luby, S, C Mendoza, B Keswick, T Chiller, and R Hoekstra. 2008. “Difficulties in bringing point-of-use water treatment to scale in rural Guatemala.” *American Journal of Tropical Medicine and Hygiene* 78:382–387.

Luoto, J et al. 2012. “Learning to Dislike Safe Water Products: Results from a Randomized Controlled Trial of the Effects of Direct and Peer Experience on Willingness to Pay.” *Environment, Science and Technology* 46(11):6244–6251.

Luoto, Jill et al. 2011. “What Point-of-Use Water Treatment Products do Consumers Use? Evidence from a Randomized Controlled Trial among the Urban Poor in Bangladesh.” *PLoS ONE* 6(10). Retrieved (http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0026132).

Makutsa, P et al. 2001. “Challenges in Implementing a Point-of-Use Water Quality Intervention in Rural Kenya.” *American Journal of Public Health* 1(10):1571–1573.

Meyerowitz, B, and S Chaiken. 1987. “The effect of message framing on breast self-examination attitudes, intentions, and behavior.” *Journal of Personality and Social Psychology* 52 (3):500–510.

Mintz, E, J Bartram, P Lochery, and M Wegelin. 2001. “Not Just a Drop in the Bucket: Expanding Access to Point-of-Use Water Treatment Systems.” *American Journal of Public Health* 91:1565–1570.

Du Preez, M et al. 2008. “Household-based ceramic water filters for the prevention of diarrhea: a randomized, controlled trial of a pilot program in Colombia.” *American Journal of Tropical Medicine and Hygiene* 79(5):790–795.

Du Preez, Martella et al. 2008. “Use of Ceramic Water Filtration in the Prevention of Diarrheal Disease: A Randomized Controlled Trial in Rural South Africa and Zimbabwe .” *The American Journal of Tropical Medicine and Hygiene* 79 (5):696–701. Retrieved (http://www.ajtmh.org/content/79/5/696.abstract).

PSI, Population Services International. 2007. *Best Practices in Social Marketing Safe Water Solution for Household Water Treatment: Lessons Learned from Population Services International Field Programs*. Bethesda, MD.

Quick, R E et al. 2002. “Diarrhea Prevention Through Household-Level Water Disinfection and Safe Storage in Zambia.” *American Journal of Tropical Medicine and Hygiene* 66(5):584–589.

Ravallion, Martin. 2012. “Fighting Poverty One Experiment at a Time : A Review of Abhijit Banerjee and Esther Duflo ’ s Poor Economics : A Radical Rethinking of the Way to Fight Poverty.” *Journal of Economic Literature* 103–114.

Rodrik, Dani. 2008. “The New Development Economics : We Shall Experiment , but How Shall We Learn ?” *Harvard University Faculty Research Working Paper Series*.

Rosa, G., Clasen, TF. 2009. “The global prevalence of boiling as a means of treating water in the home.” in *Safe Water for All: Harnessing the Private Sector to Reach the Underserved.*, edited by Sobsey M. Brown J, Outlaw T, Clasen TF, Jiangyong W. Washington DC: International Finance Corporation Retrieved (http://www.ifc.org/ifcext/sustainability.nsf/AttachmentsByTitle/p_SafeWaterReport/$FILE/IFC_WaterReport.pdf).

Rosenzweig, Mark R. 2012. “Thinking Small : A Review of Poor Economics : A Radical Rethinking by Abhijit Banerjee and Esther Duflo.” *Journal of Economic Literature* 115–127.

Rothman, A J, S C Martino, B T Bedell, J B Detweiler, and P Salovey. 1999. “The Systematic Influence of Gain- and Loss-Framed Messages on Interest in and Use of Different Types of Health Behavior.” *Personality and Social Psychology Bulletin* 25(11):1355–1369.

Smith, J K, A S Gerber, and A Orlich. 2003. “Self-Prophecy Effects and Voter Turnout: An Experimental Replication.” *Political Psychology* 24 (3):593–604.

Thaler, RH, and Cass R Sunstein. 2009. *Nudge: Improving decisions about health, wealth and happiness*. New York: Penguin.

Tversky, Amos, and Daniel Kahneman. 1981. "The framing of decisions and the psychology of choice." *Science* 211(4481):453–458.

Waddington, Hugh, Birte Snilstveit, and Howard and Lorna Fewtrell White. 2009. *Water , sanitation and hygiene interventions to combat childhood diarrhoea in developing countries*. Washington DC Retrieved (http://zunia.org/uploads/media/knowledge/3ie SR0011249554991.pdf).

Webb, Thomas, and Paschal Sheeran. 2006. "Does Changing Behavioral Intentions Engender Behavior Change? A Meta-Analysis of the Experimental Evidence." *Psychological Bulletin* 132(2):249–268.

WHO. 1997. *World Health Organization Guidelines for Drinking Water Quality*. 2nd ed. WHO, Geneva.

Figure 1: Timeline, Sample Sizes and Attrition in Both Studies

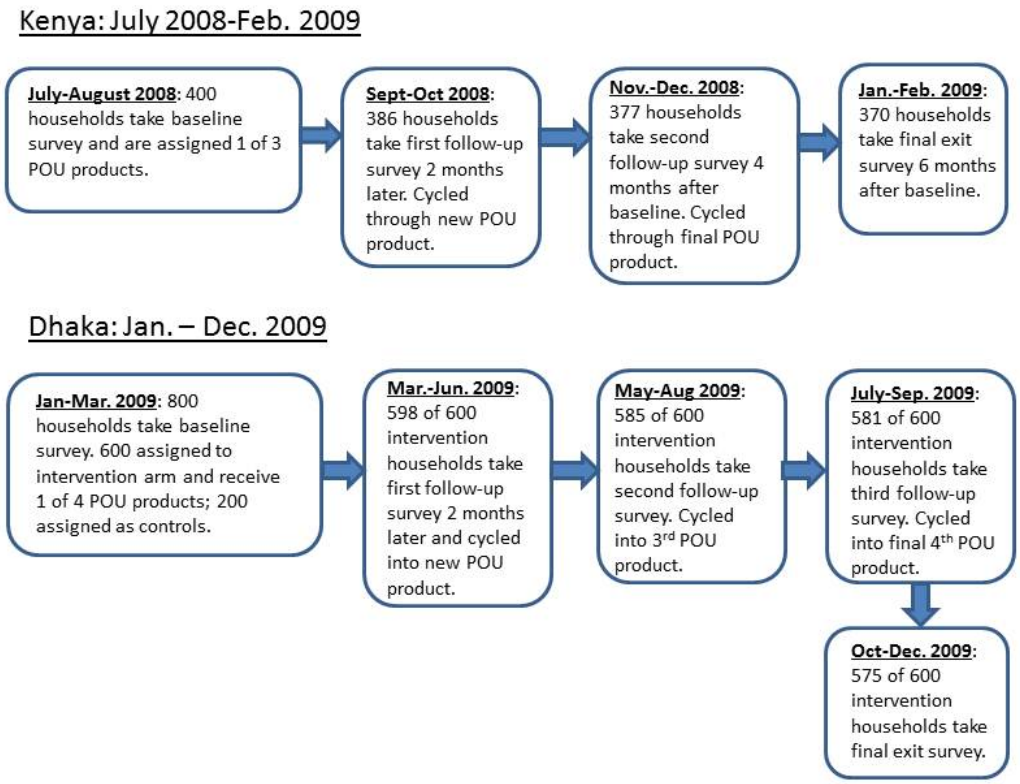
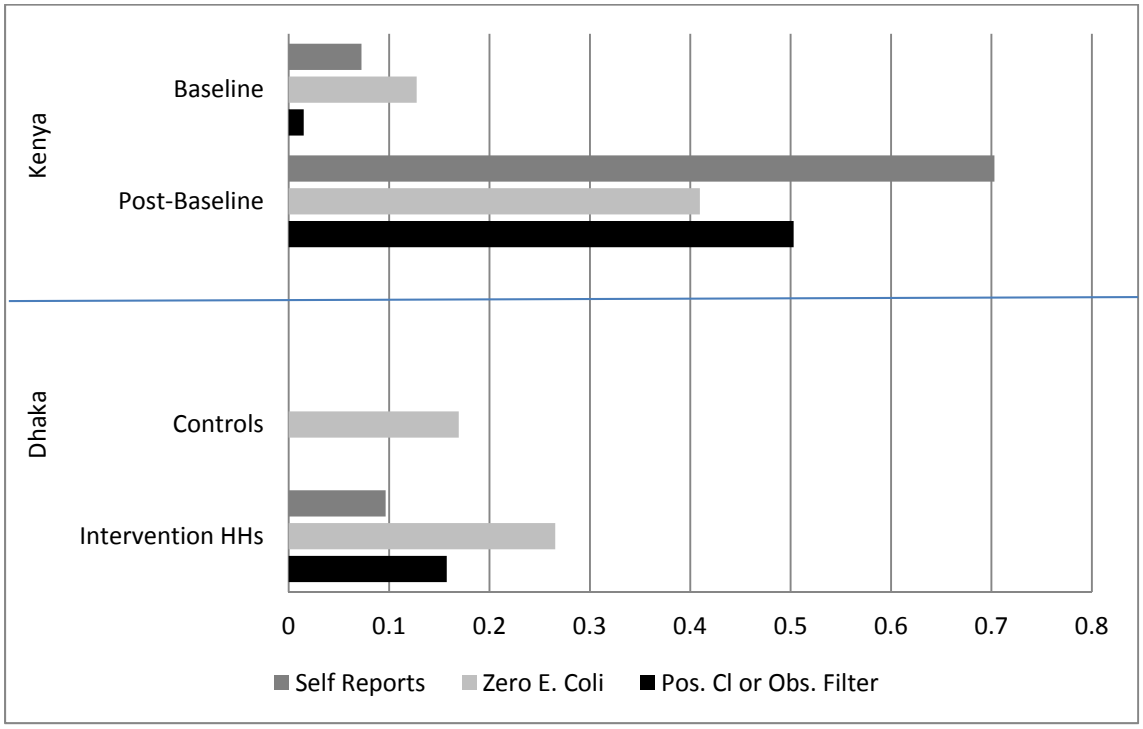


Figure 2: Base Rates Product Adoption



Rates of product usage at all two-month follow-up surveys in both settings. In Kenya all 400 originally enrolled households received free products, and thus comparison of baseline to the product averages is pre-post in design. In Dhaka, 200 of the original 800 households were designated as controls and did not receive products but had water tested every two months concurrent with intervention households. Comparing their outcomes to product averages among intervention households is an intention-to-treat comparison.

Table 1: Summary Baseline Means

Variable	Kenya			Bangladesh			
Stored Household Water Quality Variables	**Obs**	**Mean**	**s.d.**	**Obs**	**Mean**	**s.d.**	**p-value**
"Zero Risk" (No detectable *E. coli* in stored water)	377	0.14	0.34	720	0.17	0.38	0.14
Total E. coli count*		155.00			182.00		0.24
Baseline Respondent Characteristics							
Female respondent	400	0.89	0.32	800	0.98	0.13	0.00***
Married (with 1+ spouse)	400	0.90	0.30	800	0.95	0.22	0.00***
Household size	400	5.94	2.33	800	5.62	2.39	0.00***
Some secondary education or above	400	0.18	0.39	800	0.18	0.39	0.96
Illiterate adult respondent	400	0.11	0.32	800	0.43	0.50	0.00***
Respondent reports farming (formal employment) as main income source Kenya (Bangladesh)	400	0.53	0.50	800	0.18	0.38	n/a
Iron roof indicator	400	0.63	0.48	800	0.92	0.28	0.00***
Mobile phone indicator	400	0.58	0.49	800	0.61	0.49	0.24
Baseline Water and Hygiene Knowledge and Behaviors							
Respondent reports child < 5 had diarrhea in past two weeks	400	0.42	0.49	800	0.17	0.38	0.00***
Respondent lists diarrhea in top 3 diseases of concern (as top disease in Bangladesh)	400	0.56	0.50	800	0.50	0.50	0.11
Current water source is tap water	400	0.30	0.46	800	0.74	0.44	0.00***
Respondent has heard of WaterGuard	400	0.98	0.13	800	0.30	0.46	0.00***
Respondent has heard of Pur	400	0.89	0.31	800	0.08	0.28	0.00***
Respondent has heard of filter	400	0.36	0.48	800	0.54	0.50	0.00***
Respondent has heard of Aquatabs (Bangladesh only)				800	0.09	0.29	n/a
Respondent names boiling as means of diarrhea prevention	400	0.58	0.49	800	0.34	0.47	0.00***
Respondent reports boiling their water	400	0.18	0.38	800	0.34	0.47	0.80
Objectively verified soap in home at baseline visit	400	0.45	0.50	800	0.92	0.27	0.00***
Respondent self-reports POU usage at baseline	400	0.07	0.26	800	0.00	0.00	0.00***
Positive chlorine test	400	0.02	0.12	800	0.00	0.00	0.01**
Roundtrip water collection time (minutes)	400	29.90	27.15	800	12.12	14.81	0.00***
Respondent thinks source water is "safe" without treatment	400	0.42	0.49	800	0.32	0.46	0.00***

Notes: *In Kenya water quality variables are from baseline survey wave. N = 377 because 23 households had no stored drinking water on hand. In Bangladesh, water quality outcomes are based on 720 observations from 200 control households across four survey waves. All survey results (versus water quality tests) are based on complete baseline samples (including the 200 control households in the case of Dhaka). Results are identical if we restrict to the sample of those who complete the study in each country. WaterGuard in both countries is the brand name for similar liquid chlorine products. The two countries also had different ceramic filters. The last column provides p-values for unpaired t-tests of equality across experiments. ***p<.01; **P<.05; *p<.10.

Table 2: Results of Combined Marketing Treatments on POU Product Usage

	(1)	(2)	(3)	(4)
	No Detectable *E. coli*	Self-Report Usage	Product-Specific	Log(*E. coli*)
Panel A: Kenya				
1= contrast both frames (not positive frame only)	0.057	0.040	0.032	-0.227
	(0.033)*	(0.022)*	(0.034)	(0.098)**
1=commitment	0.055	0.049	0.039	-0.110
	(0.031)*	(0.023)**	(0.021)*	(0.084)
Constant	0.354	0.658	0.468	0.441
	(0.024)***	(0.019)***	(0.029)***	(0.066)***
Observations (household-wave)	1132	1120	1132	1077
Observations (households)	386	386	386	383
p-value F-test joint significance both interventions	**0.02****	**0.01****	0.144	**0.01*****
p-value F-test on sum of two interventions	**0.01*****	**0.00*****	**0.098***	**0.00*****
Panel B: Dhaka				
1= contrast both frames (not positive frame only)	0.017	0.001	0.003	-0.063
	(0.024)	(0.015)	(0.017)	(0.086)
1=commitment	0.052	0.038	0.026	-0.124
	(0.023)**	(0.014)***	(0.015)*	(0.080)
Constant	0.228	0.078	0.114	1.102
	(0.023)***	(0.014)***	(0.015)***	(0.082)***
Observations (household-wave)	2143	2339	2316	2143
Observations (households)	598	598	598	598
F-test joint significance both interventions	**0.06***	**0.02****	0.22	0.24
F-test on sum of two interventions	**0.04****	**0.08***	0.18	0.12
Panel C: Cross-equation tests of statistical significance				
Contrast frame joint test of significance	0.161	0.210	0.642	**0.050****
Commitment joint test of significance	**0.014****	**0.002*****	**0.057***	0.110
Both interventions joint test of significance	**0.004*****	**0.001*****	0.150	**0.006*****
Sum of interventions joint test of significance	**0.001*****	**0.001*****	**0.095***	**0.001*****

Standard errors in parentheses clustered at household. *p<.10, **p<.05 ***p<.01. Results from pooled OLS estimation of Equation 2 for effects of both marketing messages. Samples for both countries are restricted to those households that completed the study. All estimations include product-country fixed effects. Column 1 defines usage as treated water with no detectable *E. coli*. Column 2 indicates share of households self-reporting treatment. Column 3 is a binary outcome that equals 1 if a household had a positive chlorine test (for chemical products) or enumerator-observed filter usage, and 0 otherwise. Column 4 is a continuous measure of usage, the $\log_{10}$ of *E. coli* in a household's stored drinking water. Water with no detectable *E. coli* was assigned a $\log_{10}$(*E. coli*) value of -1. More negative values of log(*E. coli*) imply higher rates of product usage with this definition.

Table 3: Cross-Tabs of Mean Uptake under Different Marketing Message Combinations

	No Detectable *E. coli*		Self-Report Usage		Product-Specific Observed Usage	
	(1)	(2)	(3)	(4)	(5)	(6)
	Kenya	Dhaka	Kenya	Dhaka	Kenya	Dhaka
No Marketing (constant term)	0.35	0.22	0.67	0.08	0.47	0.10
	(0.03)	(0.03)	(0.03)	(0.02)	(0.03)	(0.02)
Only contrast frame	0.42	0.25	0.69	0.08	0.50	0.12
	(0.03)	(0.02)	(0.03)	(0.01)	(0.03)	(0.01)
Only commitment	0.42	0.28	0.70	0.12	0.51	0.16
	(0.03)	(0.03)	(0.03)	(0.02)	(0.03)	(0.02)
Both contrast frame and commitment	0.46	0.30	0.75	0.12	0.54	0.14
	(0.03)	(0.02)	(0.03)	(0.01)	(0.03)	(0.01)
Prob > F-stat(Both marketing - No Marketing)	0.006***	0.038**	0.026**	0.072*	0.101	0.083*
Post-baseline observations (household-wave)	1133	2143	1120	2339	1133	2316
Observations (households)	386	598	384	598	386	591

*p<.10, **p<.05 ***p<.01. Standard errors in parentheses take into account repeated observations of households. All outcomes are shares of household visits. Columns 1 and 2 give shares of household visits in which collected water samples produced no detectable *E. coli*; columns 3 and 4 present shares of household visits in which households self-reported "all of its drinking water is treated;" columns 5 and 6 present shares of households with either a positive chlorine test (with a chemical product) or the enumerator observed current filter usage. "No marketing" means households received only a positively framed message and no commitment treatment along with their freely provided POU product. "Only contrast frame" means households received the 'contrast' frame of a negatively and positively framed message. "Only commitment" means households received the commitment treatment and the positively framed message. "Both contrast frame and commitment" means households received both marketing treatments. The p-values are for F-tests of equality across the 'no marketing' and 'both marketing' conditions.

Table 4: Ordered Logit Results on Cumulative Usage due to Marketing Treatments

	(1)	(2)	(3)	(4)	(5)	(6)
	No Detectable *E. coli*	Self-Report Usage	Product-Specific Usage	No Detectable *E. coli*	Self-Report Usage	Product-Specific Usage
	Kenya	Kenya	Kenya	Dhaka	Dhaka	Dhaka
Contrast Frame	0.341	0.239	0.213	0.144	0.048	0.065
	(0.19)*	(0.19)	(0.19)	(0.16)	(0.20)	(0.18)
Commitment	0.378	0.256	0.240	0.315	0.465	0.359
	(0.19)**	(0.19)	(0.19)	(0.15)**	(0.18)**	(0.17)**
Observations (households)	370	370	370	575	575	575
Prob > chi2	0.029**	0.191	0.259	0.087*	0.038**	0.098*

Robust standard errors in parentheses. *p<.10, **p<.05 ***p<.01. Results in log-odds from ordered logit estimates of cumulative effects of contrast framed message and commitment treatment on usage at all two-month follow-up survey rounds. Estimations are on the final survey wave of 370 households for Kenya and 575 households for Dhaka. Definitions of usage for each column can be found in the notes accompanying Table 2.

Table 5: Placebo Regressions - Results of Marketing Treatments on Untreated (Source) Water Quality in Kenya

	(1)	(2)
	No Detectable *E. coli*	Log(*E. coli*)
Panel A: Kenya		
1= contrast both frames (not positive frame only)	-0.025	-0.004
	(0.02)	(0.094)
1=commitment	-0.003	0.027
	(0.02)	(0.094)
Constant	0.135	1.336
	(0.02)***	(0.08)***
Observations (household-wave)	1133	977
Observations (households)	386	381
F-test joint significance both interventions	0.577	0.958
F-test on sum of two interventions	0.374	0.855

Standard errors in parentheses clustered at household. *p<.10, **p<.05 ***p<.01. Results from OLS estimation of Equation 2 for effects of both marketing messages on un-treated water quality for Kenyan households (see text).

Appendix

The POU Products in Both Countries

WaterGuard

The U.S. Centers for Disease Control and Prevention (CDC), together with the Pan American Health Organization, developed the Safe Water System (SWS) in response to the need for an inexpensive and simple intervention that delivers clean drinking water to the poor in developing countries. The SWS involves three components. One, contaminated water is treated with a sodium-hypochlorite solution (marketed in Kenya and Dhaka as WaterGuard). Two, water should be stored in a proper manner to prevent recontamination. This generally means containers with a narrow mouth, lid, and spigot, so that people's hands do not come into direct contact with the water. Three, educational and behavior change techniques should be implemented to establish a link between contaminated water and disease and to encourage improved personal hygiene and water storage practices as well as regular treatment of water. The SWS arguably has been the most widely implemented POU measure in developing countries and the subject of the greatest number of randomized controlled studies to establish its efficacy in combating diarrheal illness. These studies largely agree on SWS's ability to reduce overall diarrheal incidence as well as the incidence in children less than five years old (Crump et al. 2005; S. Luby et al. 2001; Makutsa et al. 2001; R E Quick et al. 2002). Moreover, an overview study of the cost-effectiveness of various interventions found SWS to be the most cost-effective intervention aimed at improving water and sanitation (Hutton and Haller 2004). SWS is also found to be appropriate and effective in a variety of settings with a variety of source water qualities (Mintz et al. 2001).

To use WaterGuard: Add one capful of solution into 20 L of water (the standard jerrycan size in Kenya). If the water is turbid, add two capfuls. Stir the water briefly and then let rest for 30 minutes before drinking.

In conjunction with a free bottle of WaterGuard, our Kenyan study provided 20 L buckets with covers and taps. This was done to prevent recontamination and thereby make this product more directly comparable to the filter used in Kenya, which includes safe storage in its product design.

Pur

Manufactured by Procter & Gamble, PUR is a flocculant-disinfectant powder produced in single-use sachets that clean 10 L of water at a time. Since its introduction in 2003, a growing number of field trials have documented its efficacy in cleaning water and reducing diarrheal morbidity in a variety of settings (Chiller et al. 2006; Crump et al. 2005). PUR is particularly effective at cleaning turbid water; its flocculant powder is capable of turning brown water clear.

Using PUR involves considerably more steps than using the other two POU measures. One needs to add one sachet to a bucket containing 10 L of water, stir the water briskly for five minutes, wait five more minutes to let the impurities settle, use a cotton cloth to filter the water into a separate storage vessel, let that set for 20 minutes until it is clean, and properly dispose of the residual impurities.

Together with a two-month supply of PUR, our Kenyan study provided two buckets with covers—one with a tap for safe storage and one without a tap for the preparation process. Again, this was done to allow PUR homes to have safe storage and thereby make this product more directly comparable to the filter that was used in the Kenyan study. Buckets were not distributed along with the product in Dhaka.

Ceramic Filters (Kenya)

A variety of field studies have documented the efficacy of ceramic water filters in reducing diarrheal incidence in a variety of developing country settings (T Clasen et al. 2005; Lantagne 2001). However, the efficacy of filters has been found to be lessened in settings with turbid source waters, which slows the filtration process (J Brown and M. Sobsey 2006).

There are currently many different styles of ceramic filter designed to treat water at the household level. For this study, we used Stefani ceramic candle-shaped water filters. The filter design consists of two 20 L buckets stacked one on top of the other. Untreated water is poured into the top bucket and gravity causes the water to flow through the Stefani porous ceramic filters into the bottom bucket, which then dispenses cleaned water through a tap. Thus, the use of a filter involves just one step for households—filling it with water.

Siphon Filter (Dhaka)

The CrystalPur filter is distinct from the more common gravity-driven filters because it utilizes a siphon-driven pressure gradient to draw water through the filter element. (The manufacturer of the product provided third-party laboratory results from Waterlaboratorium Noord (May 2008) indicating >5 $\log_{10}$ reduction of *E. coli* in two tested filters. Our team replicated similar *E. coli* reduction in our own laboratory tests.) The siphon filter can sit in the stored water container if users are willing to wait to draw water through the filter when they want to drink or use the filtered water. Alternatively (and more commonly in our setting) users can filter water from a transport container into a storage container. The filter has a production rate of up to 4 L/hr, declining to 1 L/hr as the water level in the vessel declines and as solids accumulates within the filter. Users have several means maintenance options to restore filter flow after it accumulates solids including cleaning the filter's sleeve, backwashing, and scrubbing the ceramic surface with an abrasive.

Figure A1. Tested POU products in Dhaka.

Aquatabs (A), the CrystalPur siphon filter (B), the PUR Purifier of Water flocculant/disinfectant mixture (C), and dilute hypochlorite solution branded as Water Guard (D).

Water Testing Procedures in Kenya

We tested (a) source waters, (b) stored untreated water, and (c) stored treated water for turbidity, *E. coli*, and free chlorine residual (in treated water samples in which either PUR or WaterGuard had been used). Turbidity testing was performed using a portable turbidimeter (Model 2100P, Hach Company, Loveland, CO). In heavily contaminated waters, *E. coli* measurement was conducted using Petrifilm *E. coli*/coliform count plates (3M, St. Paul, MN). In samples anticipated to have lower (<3000 CFU/100 ml) concentrations, we used the Colilert Quantitray-2000 assay (IDEXX Laboratories, Westbrook, ME). Free-chlorine residual was measured using othotolidine (OTO) test kits (ILP/Swimline, Edgewood, NY).

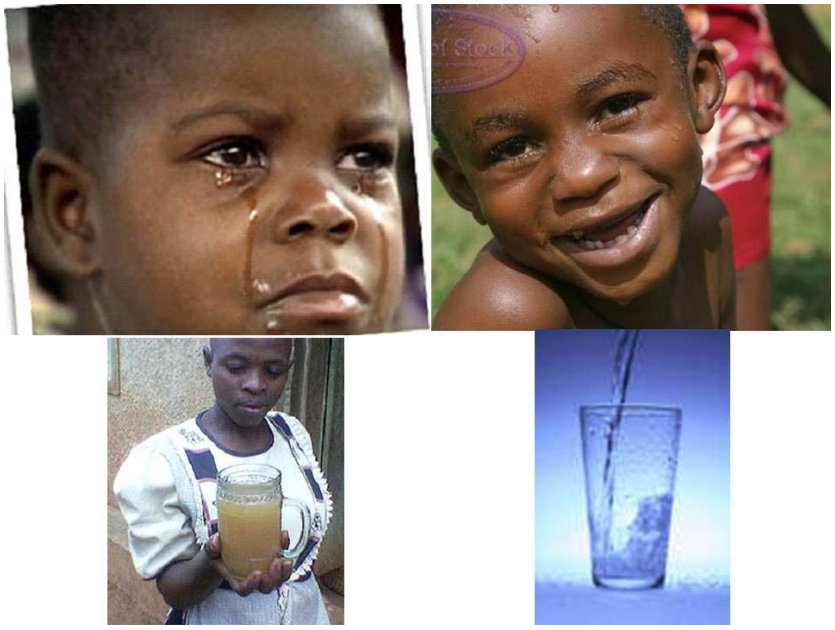
Kenyan Marketing Visual Aids

Dhaka marketing materials were qualitatively similar but featured people of South Asian descent in the images.

Figure A1: Positive Frame

A rough English translation of the corresponding verbal script read aloud to respondents with this set of images is: “By using one of these safe water products, you will be more likely to have clean, safe drinking water, which can help to keep your child[ren] happy and healthy. Use of a safe water product can make it more likely that your days will be healthy, when you can get your important tasks done. And, treating your water makes it more likely that your children will be healthy so they can grow, attend school, and learn. A safe water product can help you to achieve a healthier life. A healthier life is a happier life."

Figure A2: Contrast (Positive + Negative) Frame



A rough English translation of the corresponding verbal script read aloud to respondents with this set of images is: “Here is a picture of a sad, sick boy from drinking dirty water like we have around here. Here is a picture of a happy, healthy boy. His mother is doing many things to ensure he is having a healthy life and is happy. You also have the strength and the ability to bring such happiness to your children if you provide them with treated water. Use of a safe water product can make it more likely that your days will be healthy, when you can get your important tasks done. And, treating your water makes it more likely that your children will be healthy so they can grow, attend school, and learn. A safe water product can help you to achieve a healthier life. A healthier life is a happier life.”

Figure A3: Baseline Commitment Poster

Households assigned to receive the "commitment treatment" were given this poster at the end of the baseline visit. They were also read an additional verbal script whose approximate English translation is: "Before I leave, I would just like to ask you one more thing. You've told me that your child[ren]'s health is important to you and that your child has suffered diarrhea before. Do you want to avoid diarrhea in the future? (WAIT FOR RESPONSE) Do you believe treating your water is important to make it safe to drink? (WAIT FOR RESPONSE) Do you intend to use your safe water product every day for all your children's drinking water to keep them healthy? (WAIT FOR RESPONSE) Will you please say to me, "I will use this safe water product to keep my family's drinking water safe." Finally, as an additional way to remind you to treat your water with your safe water product every day, I'm hoping you will accept this small poster as a gift. Will you hang this poster on the wall in your home to remind you to treat your water every day? Thank you." ENUMERATOR GIVE POSTER TO RESPONDENT.

Figure A4: Sample Personalized Commitment & Reminder Poster

Sample "personalized" commitment poster from Kenya distributed to households that received "commitment treatment" at follow-up 1 interview.

Balanced Treatment Groups

Our studies included many types of randomizations; all the individual-level randomizations were implemented orthogonally to each other. We feel confident that our individual-level randomizations were effective in both countries. In Dhaka, the chi-squared test p-value was 0.67 in a probit regression that predicts treatment versus control as a function of baseline literacy, household size, native Urdu speaker, type of source water, and respondent age and gender. In Kenya, of the 55 baseline household descriptive variables compared across each individual-level randomized treatment assignment of the first product assigned, the framing message received, and the commitment treatment received, 54 (98%) balance (p-value > .1 for F-test of equality of means) across first product assigned, 52 (92%) balance (p-value > .1 on t-test for equality of means) across frames, and 53 (96%) balance (p-value > .1 on t-test for equality of means) across commitment treatment status. Furthermore, all baseline variables describing a household's water quality and collection habits balance across the individual-level randomizations.

Table A1: Marketing Treatments with Full Sets of Controls, Both Countries

	Kenya	Bangladesh
	(1)	(2)
	No Detectable *E. coli*	No Detectable *E. coli*
1= contrast both frames (not positive frame only)	0.049	0.010
	(0.029)*	(0.024)
1=commitment	0.056	0.048
	(0.029)*	(0.022)**
Female dummy	0.002	0.082
	(0.048)	(0.094)
Mobile phone dummy	0.003	0.049
	(0.030)	(0.023)**
Permanent roof dummy	0.006	-0.028
	(0.032)	(0.041)
Household size	-0.019	-0.015
	(0.006)***	(0.004)***
Age	0.001	0.002
	(0.001)	(0.001)**
Some secondary education dummy	0.091	0.095
	(0.039)**	(0.030)***
Constant	0.434	0.131
	(0.071)***	(0.110)
Observations	1133	2143

Standard errors in parentheses clustered at household. *p<.10, **p<.05 ***p<.01. Results from estimation of Equation 2 for effects of both marketing messages with full set of baseline control variables (see text).

Table A2: Marketing Treatments Interacted, Both Countries

	(1)	(2)	(3)	(4)
	No Detectable *E. coli*	Self-Report Usage	Product-Specific	Log(*E. coli*)
Panel A: Kenya				
1= contrast both frames (not positive frame only)	0.069	0.025	0.026	-0.259
	(0.043)	(0.043)	(0.041)	(0.128)**
1=commitment	0.068	0.035	0.035	-0.142
	(0.042)*	(0.041)	(0.042)	(0.130)
1=contrast frame and commitment	-0.025	-0.030	0.009	0.064
	(0.059)	(0.057)	(0.058)	(0.176)
Constant	0.347	0.665	0.470	0.457
	(0.029)***	(0.030)***	(0.031)***	(0.096)***
Observations (household-wave)	1133	1120	1133	1077
Observations (households)	386	384	386	383
F-test on sum of two interventions	**0.060***	0.412	0.404	**0.079***
Panel B: Dhaka				
1= contrast both frames (not positive frame only)	0.022	0.001	0.027	-0.039
	(0.033)	(0.019)	(0.022)	(0.121)
1=commitment	0.058	0.040	0.058	-0.093
	(0.04)	(0.024)*	(0.027)**	(0.141)
1=contrast frame and commitment	-0.010	-0.003	-0.048	-0.046
	(0.049)	(0.030)	(0.033)	(0.172)
Constant	0.225	0.077	0.097	1.086
	(0.027)***	(0.016)***	(0.018)***	(0.101)***
Observations (household-wave)	2143	2339	2316	2143
Observations (households)	598	598	591	598
F-test on sum of two interventions	0.213	0.283	**0.048****	0.574

Standard errors in parentheses clustered at household for both countries. *p<.10, **p<.05 ***p<.01.

Table A3: Marketing Results Interacted by Product, Both Countries

	(1)	(2)	(3)	(4)
	No Detectable *E. coli*	Self-Report Usage	No Detectable *E. coli*	Self-Report Usage
	Kenya	Kenya	Dhaka	Dhaka
Contrast both frames (Base effect for WaterGuard)	0.091	0.065	0.026	-0.044
	(0.051)*	(0.04)	(0.04)	(0.03)
Contrast both frames X Pur	-0.096	-0.036	-0.046	0.066
	(0.07)	(0.06)	(0.05)	(0.034)*
Contrast both frames X Filter	-0.007	-0.038	-0.029	0.042
	(0.07)	(0.06)	(0.05)	(0.04)
Contrast both frames X Aquatabs			0.034	0.067
			(0.05)	(0.038)*
Commitment (Base effect for WaterGuard)	0.101	0.093	0.060	0.060
	(0.051)**	(0.044)**	(0.04)	(0.028)**
Commitment X Pur	-0.078	-0.063	-0.049	-0.063
	(0.07)	(0.06)	(0.05)	(0.031)**
Commitment X Filter	-0.056	-0.071	0.034	0.023
	(0.07)	(0.06)	(0.05)	(0.04)
Commitment X Aquatabs			-0.015	-0.045
			(0.05)	(0.04)
Constant (Base effect for WaterGuard)	0.412	0.680	0.262	0.126
	(0.044)***	(0.041)***	(0.040)***	(0.029)***
Pur (relative to WaterGuard)	-0.087	-0.094	-0.012	-0.104
	(0.06)	(0.06)	(0.05)	(0.031)***
Filter (relative to WaterGuard)	-0.090	0.029	-0.068	-0.028
	(0.06)	(0.05)	(0.05)	(0.03)
Aquatabs (relative to WaterGuard)			-0.049	-0.059
			(0.05)	(0.04)
Observations (household-wave)	1133	1120	2120	2316
Observations (households)	386	384	590	591
F-test joint significance of contrast frame across products	0.120	0.461	0.61	0.272
F-test joint significance of commitment across products	0.199	0.193	**0.095***	**0.022****

Standard errors in parentheses clustered at household for both countries. *p<.10, **p<.05 ***p<.01.

Background information for calculated health effects

Appendix Table S1 of Fischer Walker et al. (2012) estimates that 24.8% of deaths of 1-59 month olds in Kenya are due to diarrhea. Meanwhile, the 2009 Kenya DHS survey estimates under five mortality in Nyanza province at 149 per 1,000 live births, and neonatal mortality (deaths between birth and one month) is estimated to be 39 per 1,000 live births (DHS 2009). We therefore estimate a mortality rate for 1-59 month olds of 110 per 1,000 live births that survive up to one month. Thus, in Nyanza, we estimate that 27.3 1-59 month olds per 1,000 will die due to diarrhea before reaching their fifth birthday. If we assume that these deaths are uniformly distributed across the five years, then in any given year, approximately 5.5 1-59 month olds per 1,000 live births that survive up to one month will die with diarrhea as the cause.

The general fertility rate (GFR) is 179 births per 1,000 women aged 15-44 each year in all of rural Kenya (DHS 2009). We assume a constant GFR to imply that there are 895 children under five per 1,000 women aged 15-44. This is likely a lower bound for Nyanza since it is the GFR for all of rural Kenya and Nyanza has higher fertility than average. Since neonatal mortality is 39 per 1,000 live births, this implies there are roughly 860 children aged 1-59 months per 1,000 households.

We take the sole estimate of POU products' effects on child mortality of a 42% reduction (Crump et al. 2005), and use our experiment's findings that that the marketing treatments in tandem increase rates of safe water (*defined as no detectable E. coli*) from roughly 35% to roughly 45% (Table 2).

If we assume each household has one woman aged 15-44 (likely a lower bound due to polygamy in this part of Kenya, as well as mothers and daughters living together more than offsetting the numbers of households without a woman in this age range), then for every 1,000 households we expect 860 children ages 1-59 months. We expect that 4.7 of those children aged 1-59 months will die from diarrhea in a given year.

In Dhaka, we perform a similar calculation but have lower initial fertility and child mortality, and usage of free POU products increases from roughly 22% to 30% with marketing. The 2011 Bangladesh DHS estimates (for urban areas) child mortality at 53 per 1,000 live births, and neonatal mortality is estimated to be 32 per 1,000. This implies a mortality rate of 21 per 1,000 1-59 month olds. The same Appendix Table S1 of Fischer Walker et al. (2012) estimates that 27.6% of deaths of 1-59 month olds in Bangladesh are due to diarrhea. This implies 5.8 per 1,000 1-59 month olds will die from diarrhea before their fifth birthday. We assume these deaths are once again uniformly distributed across years, so that 1.2 deaths per 1,000 1-59 month olds in a given year.

Also in the Bangladesh DHS, the GFR is estimated at 79 per 1,000 for urban women. Taking neonatal mortality into account and once again assuming constant fertility rates across years, this suggests 382 1-59 month olds per 1,000 women aged 15-44 in Dhaka. If we again assume each household has one woman aged 15-44, a policy of free provision of chlorine for one year to 1,000 households without marketing would still result in .4 deaths (versus .44 deaths with no policy). With marketing, this number is .38 deaths. This suggests free provision saves 1.2

DALYs without marketing and another .45 with marketing, at a cost of $3337 per DALY without marketing or $2455 with marketing. Per capita GDP in Bangladesh in 2008 was estimated to be $509, suggesting this policy is not cost-effective.

Our estimated benefits for both countries may be too low because we include neither results on morbidity, nor medical costs saved (for both households and provider governments or NGOs; Bangladesh's DHS suggests one out of four diarrheal episodes of a child results in a doctor's visit) nor the time saved by avoiding illness, nor any episodes of older children and adult diarrhea averted.

PAYING FOR PERFORMANCE WHEN HEALTH CARE PRODUCTION IS MULTI-DIMENSIONAL: THE IMPACT OF RWANDA'S NATIONAL PROGRAM ON REWARDED SERVICES, MULTITASKING AND HEALTH OUTCOMES

Tisamarie B. Sherry
Harvard Medical School
tisamarie_sherry@hms.harvard.edu

Sebastian Bauhoff
RAND Corporation
sbauhoff@rand.org

Manoj Mohanan
Duke University
manoj.mohanan@duke.edu

January 2013

Abstract: Performance-based contracting is particularly challenging in settings such as health care, where joint production, multiple agents, and market failures compound the critical contracting concern of multitasking. We analyze impacts of Rwanda's national pay-for-performance (P4P) program using two waves of data from the Rwanda Demographic and Health Surveys collected before and after the randomized roll-out. We find that P4P improved some rewarded and unrewarded services, but had no impact on health outcomes. We find no evidence of multitasking, and find mixed effects of the program by baseline levels of facility quality, with most improvements seen in the middle quality tier.

We are grateful to William Beardslee, Katherine Donato, Richard Frank, Martin Hoff, Mireille Jacobson, Thomas McGuire, Macros Vera Hernandez, and Alan Zaslavsky, and participants at Harvard HCP Seminar, Harvard Health Economics Workshop, iHEA 2011, U. Washington/ IHME, VCU Health Policy Seminar, the World Bank, and Yale HCP Colloquium for comments and suggestions.

1. Introduction

Contracts that link payments to performance can align the incentives of agents with the objectives of principals (Holmstrom and Milgrom 1991). However, since not all effort is contractible and contingent contracts are incomplete, agents may engage in multi-tasking (neglecting services that are unrewarded or rewarded less generously), or skimp on unobservable quality. Further, when multiple processes are rewarded and there are commonalities in the production of these processes, performance-based contracting yields ambiguous effects on both rewarded and unrewarded activities (Sherry 2012). This result is due to the potentially opposing own- and cross-price effects in a multidimensional Holmstrom-Milgrom production function.

These issues are pervasive in health care because of the high degree of information asymmetry between the provider and payer, the complexity of services and the critical importance of quality. Attempts to employ pay-for-performance (P4P) methods in healthcare have outpaced empirical evidence on their impact on healthcare quality, health outcomes, and on key economic mechanisms involved in performance-based contracting, such as multitasking and heterogeneity in responses to P4P. In this paper, we examine the impact of Rwanda's national P4P scheme, with a focus on multitasking by providers and heterogeneity in the effects of P4P as a function of baseline facility quality. Introduced in 2006, Rwanda's P4P program offers bonus payments based on a facility's performance on 24 indicators of service delivery in the areas of primary care, maternal health, child health, family planning and HIV/AIDS, scaled by an overall quality multiplier. The program was initially implemented in randomly selected districts, and rolled out nationally in 2008.

We leverage the randomized implementation and roll out of the program in conjunction with household data from two waves of the Rwanda Demographic and Health Surveys (DHS) that occurred before and after the roll-out. Our analysis contributes to the evidence on P4P as a targeted policy instrument with potentially complex effects on provider behavior. First, we examine the impact of the P4P program on rewarded measures. Drawing on a larger and different sample of households we replicate an earlier study of the Rwandan program's effect on rewarded services by Basinga et al. (2011). Second, we extend this analysis to examine the effect of P4P on additional rewarded measures that were not studied by Basinga et al. (2011), and on unrewarded measures to test for multitasking. Third, we assess the impact of P4P on health outcomes, thereby focusing on the ultimate policy objective of the program. Fourth, we examine heterogeneities in program impact across areas with different initial baseline quality.

We use difference-in-differences estimation to evaluate the overall impact of the program on the rewarded set of services as well as on unrewarded services, and the differential impact across three baseline quality tiers. We find that the P4P program had mixed effects on rewarded and unrewarded services. Among the rewarded services, and in line with an earlier study by Basinga et al. (2011), the program significantly increased the share of women delivering in facilities by almost 11 percentage points (against a baseline share of 29%), and the prevalence of contraception use among women who want to space or limit births by 4 percentage points (baseline 1.6%). There were no significant impacts on other rewarded services, such as the share of pregnant women receiving tetanus vaccination, completing 4 prenatal care visits, or receiving malaria prophylaxis, or on the rates of childhood immunizations and curative care visits among young children. We discuss possible explanations, including low real incentives (rewards net of

costs), increasing marginal costs and features of the health care production function.

To assess multitasking, we also examine the impact of P4P on measures of prenatal care quality and vitamin A supplementation in children, which are recommended as part of the national clinical guidelines but were not rewarded under P4P. Among these unrewarded services, P4P increased urinalysis by five percentage points against a baseline of 9% and increased iron supplementation by 8 percentage points (baseline of 30%), but had no significant impact on the other prenatal care measures or child vitamin A supplementation. This suggests that providers did not skimp on the key activities that are individually unrewarded but are part of the broader, rewarded service category of prenatal care visits. We also find that P4P had no significant impact on health outcomes that are available in the DHS data, such as the tested prevalence of anemia in mothers and children, and reported breastfeeding.

Finally, we find evidence of heterogeneity in the impacts of P4P, with facilities in the middle of the baseline quality distribution generally showing larger improvements in both rewarded and unrewarded services.

The remainder of the paper proceeds as follows: in Section 2 we discus intended and potential unintended consequences of P4P programs, and the evidence to date on the impacts of these programs. Section 3 describes Rwanda's P4P program and its randomized roll-out, and motivates the hypotheses we aim to test in this paper. Section 4 provides details on the data. Sections 5 and 6 describe our empirical strategies and present our results, and Section 7 discusses our findings in light of the theoretical predictions.

2. Evidence on the Intended and Unintended Consequences of Pay-for-Performance

Performance-based payments can affect both rewarded and unrewarded activities (Holmstrom and Milgrom 1991). When two activities are substitutes in an agent's cost function (i.e. effort devoted to one activity increases the marginal cost of effort devoted to the other activity) and only one of these activities is rewarded with incentive pay, then the output of the rewarded activity is expected to increase at the expense of the unrewarded activity. This multitasking problem arises when only a subset of providers' activities are rewarded – a setup common in P4P programs in health care (Prendergast 1999).

However, the classic intuition of multitasking can break down when P4P rewards multiple targets. In the presence of multiple targets, own- and cross-price effects from a Holmstrom-Milgrom production function work in different directions to yield an ambiguous net effect on both the rewarded and unrewarded activities (Sherry 2012).

Performance-based contracting is particularly challenging in complex settings such as in health care, due to significant information asymmetries, conflicting incentives, multiple inputs and actors, and joint production by both providers and patients (Arrow 1963). In contrast to pay-for-performance efforts in sectors such as education, where financial incentives for teachers are often tied to outcomes such as student performance (Glewwe, Ilias et al. 2010; Muralidharan and Sundararaman 2011), most P4P efforts in health care reward providers based on process measures, which are inputs into the production of health. There are very few instances of outcome-contingent contracts in health - see Leonard (2003) for a notable exception -

possibly because of pervasive incomplete contracts and verifiability problems in measurement of health outcomes.

Another key feature of health care settings that complicates the potential impacts of P4P is joint production or "commonality" (Glazer, McGuire et al. 2008; Mullen, Frank et al. 2010). Joint production occurs when a common input affects the output of multiple services. Examples that are relevant to the Rwandan setting include the availability of clean syringes as a common input in both childhood immunizations and prenatal tetanus immunizations; or HIV voluntary counseling and testing services as a common input in both HIV treatment and the dispensation of modern contraceptive methods that limit the spread of sexually transmitted infections, e.g., condoms. As a result of joint production, P4P will not necessarily decrease the supply of unrewarded services: the overall impact will depend both on the degree of multitasking and the existence of commonalities in production (Mullen, Frank et al. 2010; Sherry 2012). With regard to rewarded services, P4P will be more likely to increase the output of services that (1) earn a higher bonus payment relative to their production cost (McGuire and Pauly 1991); and (2) are jointly produced with other highly rewarded services.

Evidence on the impact of targeted performance incentives on physician behavior in the US is mixed, and many studies of P4P have been observational and have relatively small samples. Most studies have found no or at best modest improvements in quality, but also no disruptions in care (Rosenthal and Frank 2006; Christianson, Leatherman et al. 2008; Mullen, Frank et al. 2010). Possible explanations for P4P's muted impact in the US include the relatively small size of performance incentives, and the small share of a given provider's panel that is enrolled in a health plan implementing P4P (Rosenthal and Frank 2006). Yet studies of P4P programs

implemented by larger payers offering larger bonuses have also found only modest impacts on quality (Rosenthal, Frank et al. 2005; Glickman, Ou et al. 2007; Lindenauer, Remus et al. 2007; Mullen, Frank et al. 2010; Werner, Kolstad et al. 2011). Even in the context of single-payer settings like Ontario, Canada, P4P incentives have led to only modest improvements in some rewarded services and no improvements in others (Li, Hurley et al. 2011). Similarly, the size of the financial incentive also does not appear to be the primary explanation for the absence of impact. The UK's P4P program for family practitioners, for example, has been associated with only modest short-term improvements in quality indicators (Campbell, Reeves et al. 2007; Campbell, Reeves et al. 2009) despite its comprehensive approach and substantial financial incentives: under the P4P contract, high-performing family practitioners stood to increase their income by up to 25% (Doran, Fullwood et al. 2006). There is emerging evidence on P4P from middle- and low-income countries. An evaluation of a P4P program in the Philippines found improvements in provider knowledge as measured by vignettes (Peabody, Shimkhada et al. 2011). Previous studies of Rwanda's national P4P program showed improvements in rewarded services (Basinga, Gertler et al. 2011) and positive impacts on certain health outcomes (Gertler and Vermeesch 2012). Olken, Onishi et al. (2011) also find positive health impacts of performance based financing for villages in Indonesia and improvements in healthcare workers' labor.

Only a limited number of studies have explicitly tested for unintended consequences of P4P such as multitasking, or for the impact of P4P on health outcomes, and their findings have also been mixed (Glickman, Ou et al. 2007; Mullen, Frank et al. 2010). The UK's national P4P program, for example, witnessed improvements in some unpaid indicators, but declines in other unpaid measures (Steel, Maisey et al. 2007; Campbell, Reeves et al. 2009; Sutton, Elder et al. 2010). Finally, there are concerns that the incentive

structures used by many P4P schemes may discourage improvement by the lowest and/or highest performers (Rosenthal, Frank et al. 2005; Petersen, Woodard et al. 2006; Rosenthal and Dudley 2007). High and low performing facilities are likely to have different marginal costs of production, so their responses to different types of incentives could vary considerably. Understanding the relationship between baseline performance and the impact of P4P is therefore especially important as these initiatives are scaled up.

3. Rwanda's National P4P Program

Rwanda's national P4P scheme was launched in May 2006, with the goal of improving the quality of care provided in the country's 38 district hospitals and 420 district health facilities (Basinga, Gertler et al. 2011). We focus on the program implemented in district health facilities (hereafter "facilities"), which provide the majority of primary care services (Hoff 2010).

Prior to 2006, facility budgets were input-based and were determined prospectively based upon the number of patients each facility was expected to serve (Hoff, 2010). The P4P program supplemented these budgets with bonus payments to the facilities based on their performance on 24 output and quality indicators. The unit payments for rewarded indicators are shown in Appendix Table A1. The size of the payments varies considerably across services, from US $0.09 for each first-time prenatal care visit to US $8.93 for HIV testing of each exposed child. In general, institutional deliveries and HIV/AIDS care are most generously rewarded services under the P4P scheme. P4P bonus payments are disbursed to facilities quarterly and the facilities may allocate the payments at their discretion. In the first two years, the bonuses increased facilities' budgets by 22% on average. Overall, 77% of the bonuses were used to increase compensation, resulting in a 38% increase

in staff salaries (Basinga, Gertler et al. 2011). The remaining 23% was typically spent on infrastructure or health care inputs.

A unique feature of Rwanda's P4P program is that bonus payments are scaled by a quality multiplier *M* reflecting the facility's overall performance. *M* may take any value between 0 and 1, where higher values indicate higher overall facility quality. Overall facility performance is assessed on a broad range of service categories such as infrastructure, personnel, supplies, and adherence to clinical practice guidelines. The service quality categories are listed in Appendix Table A1. Quality scores for each category are determined by district health officials based on unannounced quarterly assessments. The overall quality multiplier is constructed as a weighted average of the quality scores across all service categories.

The overall bonus payment formula therefore scales the total payment to each facility in a payment period by the value of the quality multiplier *M* assigned to each facility for that period:

$$Bonus = M \times \left(\sum_{n=1}^{24} p_n \, r_n \right) \qquad (1)$$

Where r_n is the number of times rewarded indicator n is met and p_n is the corresponding unit payment for rewarded indicator n, and M is the value of the quality multiplier.

This divergence between the explicitly targeted services and the services that are part of the broader quality multiplier can potentially mitigate the standard multitasking problem. A number of the service categories covered in the quality multiplier – for example, curative care, delivery, prenatal care, family planning, immunization and HIV services – overlap with

specific services that are explicitly rewarded with P4P bonus payments. Other service categories in the quality multiplier, however, such as general administration, cleanliness, laboratory services and pharmacy management, are not explicitly targeted by P4P. Since these services are inputs to a broad range of services, including them in the multiplier might contain the effect of multitasking.

The quality multiplier also creates varying incentives for facilities with different levels of overall quality. Initially, higher-performing facilities with quality multipliers closer to 1 face the highest marginal incentive for each unit of rewarded service. Lower-performing facilities, on the other hand, face strong incentives to increase their quality multiplier to eventually earn larger overall P4P bonus payments. If the costs of production of healthcare are not increasing in quality, we would expect facilities with higher baseline performance to achieve larger absolute increases in services rewarded under P4P and smaller increases in unrewarded services, relative to facilities with lower baseline performance. However, without knowledge of the facilities' production function, it is difficult to predict the relative influence of the incentive payments.

3.1 The Randomized Roll-Out of P4P

The Rwandan Ministry of Health instituted a phased and quasi-experimental roll-out of the national P4P program to facilitate evaluation.[1] During Phase 0 (2002 – 2006) several non-governmental organizations implemented P4P programs in 11 of Rwanda's 30 districts. Basinga et al. (2011) describe the block-randomization of the remaining 19 districts based on rainfall and Census population density and livelihood measures. The districts were randomized into 12 treatment districts and 7 control districts in

[1] This paragraph draws on the description of Rwanda's P4P in Basinga et al (2011).

Phase 1. Together with the 11 districts that had implemented pilot programs in Phase 0, the 12 treatment districts adopted the national P4P scheme between June and October 2006. In Phase 2, from April through May 2008, the program was rolled out to the 7 control districts. To isolate the effect of giving facilities performance incentives from the effect of simply increasing their available resources, during Phase 1 the Ministry of Health increased the control facilities' input-based budgets by the average bonus payment to the treatment facilities. In this study, we analye the impacts of P4P only on the randomized districts, since the Phase 0 initiatives were not randomly assigned and varied in their approaches. Our results are robust to including these early adopters in our estimations (details available upon request).

Basinga et al (2011) evaluate the impact of Rwanda's national P4P scheme on health care utilization and find that the program significantly increased the utilization of services that had higher unit payments and were easier for providers to control. This study is one of the few rigorous evaluations of P4P in a low-income country to date. However, since the study relies on a relatively small sample of 2,158 households including 1,290 women, some effect sizes were imprecisely estimated. Also, the focus of this evaluation was on program impacts on rewarded services, and it did not examine effects on unrewarded services or heterogeneities by baseline facility quality. In a more recent paper, Gertler and Vermeesch (2012) rely on follow up data collected in the same study areas in Rwanda and find significant impacts on children's weight-for-age and height-for-age, attributing these positive effects to an "increase in use and quality of prenatal and post natal care". They also find a positive impact of the program on provider performance as measured by patient exit interviews and a closer alignment between what providers know and what they do in practice.

Our analysis relies on two waves of the Demographic and Health Surveys (DHS). In addition to replicating the previous evaluation findings with a larger dataset, we examine the impact of P4P on an expanded set of rewarded measures, and focus on unintended consequences of P4P such as multitasking, heterogeneity in P4P's impacts based on initial facility quality, and the impact on a range of health outcomes, all of which are central concerns in P4P.

Given the particular structure of the Rwandan payment formula, we test hypotheses that (a) provision of rewarded services increased, as previously reported; (b) unrewarded services that share commonalities in production with rewarded services increased; (c) health outcomes associated with the rewarded services improved; and (d) improvements in rewarded and unrewarded services varied according to baseline levels of facility quality.

4. Data

The Rwanda Demographic and Health Surveys (DHS) is a large, nationally representative household survey that collects cross-sectional data on health and demography. We use the two most recent waves of the DHS for our analysis: DHS-III and DHS-IV. DHS III surveyed 11,321 women (ages 15-49) drawn from 10,272 households, between February and July of 2005. DHS IV surveyed 7,313 women drawn from 7,377 households, between December 2007 and April 2008. [2] DHS-III was collected prior to the roll-out of P4P to the treatment districts, while DHS-IV was collected 18-22 months after this initial roll-out, but prior to the expansion of P4P to control districts. Each wave collects information from female respondents about current health behaviors, any pregnancies in the preceding 5 years, and the health of

[2] While the DHS-III was a standard, full-length DHS, the DHS-IV was an interim survey with a smaller sample size and abbreviated questionnaire.

children born in the preceding 5 years. Although the DHS-IV is an interim survey with an abbreviated instrument there is substantial overlap of variables across the two waves. Since we use household reports, our measures are not subject to disproportionate over-reporting by providers in the treatment group.

Our analytical sample is a balanced panel of villages or localities containing two repeated cross-sections of households – i.e., although the same localities are visited in each wave, sample households are visited only once in either the pre- or post-intervention periods. The primary sampling units used in DHS-III and DHS-IV are enumeration areas (EAs) from Rwanda's 2002 Census and represent large villages or clusters of villages. In DHS-III, 432 EAs were sampled, while in DHS-IV a subset of 250 of these EAs were sampled. We restrict our analysis to households and individuals in the 153 EAs in phase 1 or phase 2 districts that were sampled in both DHS-III and DHS-IV.[3] Our final sample includes households in 89 EAs from the 12 treatment districts, and households in 64 EAs from the 7 control districts, allowing us to control for local-area fixed effects.

As a result of varying reference populations across the two waves of DHS, and the different recall periods across measures, the analyses include different sample sizes across different outcome variables. For example, information about institutional deliveries is available for all births within the past five years while maternal anemia is only measured for women who are currently pregnant. We use the maximum available sample for each

[3] We match EAs using geographic identifiers contained in both waves of the data. Although the geographic identifiers in the two DHS waves match up exactly, the coordinates in the public use files contain random positional error to maintain confidentiality. Urban and rural clusters are displaced by up to 2km and 5km, respectively; an additional 1 percent of rural clusters are randomly displaced by up to 10km. This could lead to misclassification if apparent control EAs are in fact located in treatment areas.

dependent variable. Table 1 summarizes the baseline characteristics of women and children in the treatment and control areas, prior to the rollout of P4P. Although our data are sampled and collected separately from those used by Basinga et al. (2011), they are broadly comparable: female household members have very low education, predominantly live with a partner, and had between 4 and 5 births at the time of the interview. Households have an average of 5-6 members, and slightly above half have insurance. As in the earlier study, we find that several of the unadjusted pre-intervention measures differ across the treatment and control groups, which may be a result of the coarse block randomization. However, no group performs clearly better than the other in the initial period and the differences are not statistically significant.

5. Average Impact of P4P

We estimate the impact of Rwanda's national P4P program on the outcomes of interest using a difference-in-differences (DD) model:

$$y = \beta_0 + \gamma \times POST + \delta \times POST \times TREAT + \beta_1 \times C + \alpha + \theta + \varepsilon \qquad (2)$$

where y is the outcome of interest; *POST* is an indicator variable for the post-intervention period[4]; *TREAT* identifies the treatment districts so that δ is the coefficient of interest; C is a vector of household and individual-level control variables; α is a vector of EA-level fixed effects; and θ is a vector of birth-year fixed effects.[5] Standard errors are clustered at the district, i.e. the level of the treatment. Because we only have G=19 clusters, we also use critical values from a t-distribution with G-2 degrees of freedom (Cameron,

[4] The time cutoff for the post indicator varies with the dependent variable. See Appendix Table A2 for details.

[5] Our covariate set compares with the most complete specification in Basinga et al (2011). Birth-year fixed effects are omitted from the models estimating the impact of P4P on contraceptive coverage.

Gelbach et al. 2008; Cohen and Dupas 2010). The critical *t*-values for the 1%, 5% and 10% significance levels are 2.90, 2.11 and 1.74, respectively.

The identification assumption is that in the absence of P4P, the evolution of outcomes in the treatment districts would have been comparable to the observed changes in the control districts. The EA-level and birth-year fixed effects eliminate area-specific unobservable factors (such as distance to facilities and local attitudes) and time-variant unobservable factors that are common to both types of districts (e.g., national policies and changes in the environment that may have influenced nutrition and health). One potential concern is the expansion of community-based health insurance over the study period; we control for health insurance expansion to eliminate confounding due to coincidental overlap with the P4P roll-out. An analytical benefit of this expansion is that the insurance schemes are coordinated and operated on the district level: until very recently, and during the period of time included in our study, households enrolled in Rwanda's community-based health insurance program were affiliated with a specific health center, and only from there could obtain referrals to other district health centers or hospitals (Lu, Chin et al. 2012). This should serve to reduce potential spillovers from individuals residing in control districts seeking care in treatment districts and vice versa (Ministry of Health of Rwanda 2010).[6]

5.1. Outcomes

[6] In a robustness check (available upon request) we also exclude EAs whose treatment assignment would be altered by this positional error (i.e. urban EAs from control districts within 2km of the boundary with a treatment district; urban EAs from treatment districts within 2 km of the boundary of a control district; rural EAs from control districts within 5km of the boundary with a treatment district; and rural EAs from treatment districts within 5 km of the boundary of a control district). Our findings remain comparable. We cannot directly test for changes in travel time and patterns, as this information is not available in the DHS.

We analyze the impact of P4P on rewarded and unrewarded services related to family planning, maternal and child health care, and health outcomes. Table 1 summarizes these outcome variables. The list of rewarded services that we focus on is a subset of the 24 measures rewarded in the P4P program that are also reported in the DHS surveys. Most of our outcome measures map very closely to the services rewarded under P4P. As a proxy measure for curative care visits, we use visits to a health care facility for treatment of an infectious illness in a child under the age of 5. We use reported use of modern contraceptive methods as proxy measure for family planning visits. In addition to the overall contraceptive coverage we also consider the subset of the women who report that they want to limit or space births; this group is "in greater need" of contraception and may be easier for providers to reach. As a proxy for contraceptive resupply we employ the prevalence of modern contraceptive methods that require regular resupplies. See Appendix Table A2 for details on the variable construction.

The unrewarded services include measures of care for children and prenatal care quality. The unrewarded measure of care for children is appropriate vitamin A supplementation for children (i.e. receipt of a vitamin A supplement in the past 6 months by children age 1-5) and seven binary prenatal care quality indicators that are available in the data: whether prenatal care was received from a trained provider (i.e. a licensed medical professional such as a doctor, nurse or midwife); whether the respondent was weighed during pregnancy; whether blood pressure was measured during pregnancy; whether a urinalysis was performed during pregnancy; whether blood was drawn during pregnancy; whether an iron supplement was given during pregnancy; and whether the respondent was explained the possible complications of pregnancy.

Health outcomes include favorable and adverse events. We include whether the child was breastfed for 6 months or more as a favorable event. Adverse outcomes include the prevalence of anemia in children and pregnant women; vision difficulties during pregnancy (any vision difficulties and adjusted night blindness)[7]; and treatment of children's infectious illness in the preceding two weeks.[8]

As Table 1 shows, observable characteristics are balanced across treatment and control districts at baseline. We include the following covariates as controls in regressions: the total number of births per woman, the wealth quintile of households, and household insurance status. These differences are a possible consequence of the block-randomization. For analyses of child outcome variables, we also control for the child's age and sex.

As we examine many outcomes we are concerned about the multiple comparisons problem. We cannot address this easily issue by constructing an index of outcomes as in Kling et al. (2007) because the various outcomes have different denominators (Kling, Liebman et al. 2007). Instead of redefining our unit of analysis as the enumeration area (which would provide a common denominator) we note that our statistically significant results for rewarded

[7] Vision changes during pregnancy at any time of the day or night may result from advanced gestational diabetes or pre-eclampsia, but because vision changes in pregnancy may also be normal, this symptom is fairly non-specific Roos, N. M., M. Wiegman, et al. (2012). "Visual disturbances in (pre)eclampsia." Obstetrical and Gynecological Survey **67**(4): 242–250.. We therefore also measure adjusted night blindness, which is a more specific symptom of pathology (vitamin A deficiency) during pregnancy (WHO, 1998). Following the DHS, we report the adjusted night blindness as the prevalence of women reporting vision changes only at night – women who also report vision changes during the day are excluded because these are not consistent with vitamin A deficiency National Institute of Statistics of Rwanda (NISR) and Macro International Inc (2005). Rwanda Demographic and Health Survey I. Macro International. Calverton, United States: .

[8] The prevalence of anemia is measured by blood tests of study children; in contrast, the presence of infectious illness in the past 2 weeks is reported by the child's mother.

services are similar to those reported by Basinga et al. (2011) on an unrelated dataset. This replication result suggest that our (few) statistically significant findings are unlikely a result of the multiple comparisons.

5.2. Results

Figure 1 summarizes the average effects of the P4P program from estimating the DD specification in equation 2. These effects are also reported in the top (Panel A) of Tables 2, 3 and 4.[9] For rewarded services, the program significantly increased the share of women delivering in facilities by 10.63 percentage points against a pre-intervention baseline of 29%. Although P4P had no significant impact on the overall rate of contraceptive use among women who were married or living with a partner, among the subset of women with the highest need for contraception (i.e. those without fertility problems who expressed a desire to space or limit births), P4P increased the prevalence of modern contraceptive use by 3.94 percentage points (baseline 1.6%). We do not find any significant impact on other rewarded services, such as appropriate tetanus vaccination and malaria prophylaxis for pregnant women, the share of women who completed 4 antenatal care visits, childhood immunizations (percent children who are fully immunized by age 1), or the share of children with a recent infectious illness who were treated at a public facility.

For outcomes that overlap with those reported by Basinga et al. (2011) our results are broadly similar to this earlier study with the exception of tetanus vaccination. Basinga et at. (2011) find statistically significant improvements in the rate of institutional deliveries (8.1 percentage points, p-value 0.017, compared to our point estimate of 10.63) and tetanus

[9] In Tables 2 – 4 we only show the main coefficients of interest. The full results are available in Appendix Tables B.1-B.6.

vaccinations (5.1 percentage points, p-value 0.057, compared to our not statistically significant estimate of 7.2). Like us, they find no statistically significant effects for the completion of four antenatal visits and childhood immunizations.

Our results indicate a mixed pattern of the impact of P4P on rewarded health services. As expected, the impact of P4P varies with the size of the reward for a particular service: deliveries are the most generously rewarded service in our data, at $4.59/delivery, while first-time family planning visits promoting contraceptive coverage also receive a relatively high P4P bonus payment of $1.83/visit. The findings suggest that variations in costs may also matter for performance, however. While there is no impact on the prevalence of modern contraception use in the general population, there is a large and statistically significant increase in contraceptive coverage for women who report wanting to limit or space births. Since this group may be motivated to actively seek out contraception, it may be less costly for providers to increase contraceptive use among these women. Consistent with this notion, we see no significant improvements in a number of rewarded services that may also be costly to provide (e.g. due to necessary repeat visits), including childhood immunizations and prenatal care visits. Finally, the marginal costs of improving a service may not be constant but may be increasing in the pre-intervention prevalence of the service. That may reconcile the lack of an effect of P4P on childhood immunizations (baseline 43%) alongside its positive effect on modern contraception use among women who want to limit or space births (baseline 1.6%), despite the fact that these services had the same reward ($1.83). These findings are also consistent with those reported by Gertler and Vermeesch (2012) who find larger effects of P4P on services that had higher rewards and a lower marginal cost of delivery.

The impact of P4P on unrewarded services is presented in Panel A of Table 3. We find that the P4P program had significant, positive impacts on urinalysis (4.92 percentage points) during prenatal care and prenatal iron supplementation (7.86 percentage points). We find no statistically significant effect on any of the other aspects of prenatal care quality or vitamin A supplementation provided to children.

Table 4 shows estimates of the impact of the P4P on health outcomes. We find that P4P had no significant impact on any of the health outcomes studied: breastfeeding for at least 6 months; night vision or any vision difficulties during pregnancy; anemia in currently pregnant women or in children; and infectious illnesses in the past 2 weeks in children under the age of 5.

An incidental finding in these analyses concerns household insurance status: having health insurance is associated with large and significant improvements in several rewarded services (institutional deliveries, percent of women completing four prenatal care visits, rate of fully immunized children and the rate of curative care visits for childhood infectious illnesses) and two unrewarded service (seeing a trained prenatal provider and child vitamin A supplementation). It is also associated with a decrease in the prevalence of child anemia. These results are only correlational and should not be interpreted as evidence of a causal relationship between insurance and health service utilization or health outcomes. Aside from insurance, the coefficients on the *POST* indicator show substantial secular improvements between 2006 and 2008, especially in the rates of childhood anemia and childhood infections, and in the use of modern contraceptive methods. While our study does not enable us to causally interpret the *POST* variable, some of the improvements in outcomes over time could be attributable to the near-quadrupling of total annual health expenditures between 2004 and 2009

(USD 130M to 531M in 2012 dollars) and an influx of external funds and government programs, including for family planning (Basinga, Gertler et al. 2011; World Health Organization 2012). The improvement in the control areas could also be due to the increase in the input-based budgets of the control facilities.

6. Heterogeneity by Pre-Intervention Quality

In the second part of our analysis, we examine how the impact of P4P varies according to an area's pre-intervention quality. Since the DHS does not identify individual facilities, it is not possible to precisely match information on health service utilization with facilities. Instead, we construct a quality score Q for each enumeration area (EA), which reflects the average quality of health care in that EA, and we use Q as a proxy for the quality of the nearest facility or facilities.

To calculate Q, for each EA we determine the percentile in the pre-roll-out period for: facility deliveries; pregnant women receiving appropriate tetanus vaccination; pregnant women receiving appropriate malaria prophylaxis; pregnant women completing four prenatal visits; children receiving vitamin A supplementation; children fully immunized; and women living with a partner who are using a modern method of contraception.. These are representative measures available in DHS data from each of the service categories that we examine in this analysis: delivery and prenatal care, child health and family planning. The quality score Q for a given EA is the average of these of these pre-intervention percentiles.

EAs are then assigned to three tiers (low, medium, and high) according to their quality score *Q*, with the bottom third of EAs in the low tier, the

middle third in the middle tier, and top third of EAs in the high tier. We then repeat our difference-in-differences analysis using the following model:

$$y = \beta_0 + \sum_{i=1}^{3} \gamma_i \times POST \times \bar{Q}_i + \sum_{i=1}^{3} \delta_i \times POST \times TREAT \times \bar{Q}_i + \beta_1 \times C + \alpha + \theta + \varepsilon \qquad (3)$$

In this specification, $\bar{Q}_i$ is an indicator variable equal to 1 if the EA in which the respondent lives belongs to the i-th tier and 0 otherwise. All other variables are the same as in the main specification in equation 2. The impact of P4P on y in tier i is therefore captured by the coefficient δ_i, and $\bar{Q}_i$ is subsumed in α, the vector of EA-level fixed effects.

6.1. Results

The lower panels (B) in Tables 2, 3 and 4 present the results of the P4P program interacted with the low, medium, and high quality tiers, following the DD specification in Equation 3. Among the rewarded health services, there are significant variations in program impacts by baseline quality for several of the services. Most of the improvements in institutional deliveries (13.30 percentage points) are concentrated in the high tier. For several other rewarded services – namely malaria prophylaxis, use of modern contraception by all married/cohabitating women and use of modern contraception requiring resupply – there are significant improvements in areas that were in the medium tier. The only service with significant improvements concentrated in low tier areas is use of modern contraception by higher-need women. Curiously, the program also led to a *reduction* in the number of prenatal visits in the low tier (-7.67 percentage points). While there are no unambiguous trends across the tiers, our findings suggest that the top quality tier achieved improvements in the most generously rewarded

service studied (i.e. institutional deliveries), but the smallest number of rewarded services overall. The median of the coefficients for rewarded services among areas in the lowest-quality tier is 0.87, compared to 5.9 for medium tier, and -0.97 for the highest-quality tier. Medium-tier facilities therefore appear to have achieved improvements across a broader array of rewarded services.

We see no significant variations in program impact by baseline quality for the percent of children fully immunized by age 1, curative care visits for children and tetanus vaccination among pregnant women.

Among unrewarded services, we see a different pattern of heterogeneity in P4P's impacts across areas with different baseline levels of quality. For one of the unrewarded prenatal care services studied - iron supplementation - improvements are concentrated in the high tier (12.23 percentage points). Conversely, in the high quality tier there are weakly significant reductions (-5.78 percentage points) in the percent of pregnant women being informed of potential complications of pregnancy. The medium tier areas show improvements in two unrewarded quality indicators: blood pressure measurement (13.09 percentage points), and urinalysis (6.97 percentage points). There are no significant improvements in unrewarded services in the low quality tier. The median of the coefficients for unrewarded services among areas in the low-quality tier is 0.26, compared to 7.72 for the medium tier, and 0.87 for the high tier. Again, improvements are concentrated in the medium-tier facilities.

The impact of P4P on health outcomes does not vary significantly by an area's baseline quality, with two exceptions. The incidence of reported infectious illnesses in children (diarrhea, cough or fever) increased in the low and medium-quality areas (10.56 and 17.01 percentage points) and decreased

in the top tier (14.74 percentage points). The share of children who were exclusively breastfed for at least 6 months increased in the medium tier (4.95 percentage points).

7. Discussion

We find that Rwanda's national P4P program had mixed effects on health care quality. The program increased the output of some rewarded services, such as institutional deliveries and contraceptive coverage for women with the highest need, while others such as tetanus vaccination, malaria prophylaxis, the number of prenatal visits, and childhood immunizations and curative care visits did not increase. We find no evidence that the output of unrewarded services decreased as a result of multitasking – indeed, P4P had a positive impact on several unrewarded aspects of prenatal care, suggesting positive spillovers possibly due to complementarities in production. We find that P4P had no significant impact on any of the health outcomes studied.

While some services that were highly rewarded increased in response to P4P, other services did not respond consistently. For example, first-time family planning visits and completion of childhood immunizations receive the same bonus payment of $1.83, but only targeted contraceptive coverage (a proxy for family planning visits) increased in response to P4P. There are several factors that might account for the discrepancy in P4P's impacts on services with similar bonus payments. First, it is possible that other public health initiatives outside of P4P might have influenced these results – indeed, the study period coincided with both a national immunization campaign and increased HIV/AIDS funding and programming, which is closely related to family planning. The positive coefficients on the POST indicator suggest the importance of such longitudinal, secular changes. To our knowledge, these

concurrent initiatives did not coincide with the roll-out of P4P – still, they might have indirectly influenced provider responses to P4P. For instance, increased HIV/AIDS programming might have sensitized patients and providers to the importance of contraceptive methods that prevent HIV transmission, creating a positive interaction with P4P. Conversely, in the case of childhood immunizations providers might have felt that there was little that they could contribute to immunization efforts above and beyond a national campaign, or the campaign's overall effects might have obscured additional but minor improvements by providers.

Second, providers respond to the net rewards, i.e. incentives net of costs, and it is possible that different services have varying and increasing marginal costs. For example, the completion of childhood immunizations may require more provider and parent effort than first-time family planning visits, since the former requires repeated interactions. We would therefore expect to see larger improvements in contraceptive coverage for the same bonus payment. The findings for general versus targeted contraceptive coverage are also consistent with this mechanism: it may be less costly to increase service provision for patients with higher demand (e.g., women wanting to limit or space births) than for the average patient. This finding also highlights the need to consider demand-side factors when designing P4P programs, and the potential for implementing concurrent demand-side interventions that can lower the marginal costs to providers. Indeed, for most of the quality measures studied, we found a significant, positive association between individual insurance coverage and the receipt of care, raising the question of how demand-side policies such as insurance expansions might interact with supply-side initiatives such as P4P.

Finally, recall that HIV/AIDS services actually received the highest rewards under Rwanda's P4P scheme, and that family planning services are

likely to be jointly produced with HIV/AIDS services given the common emphasis on safe sexual practices – indeed, Rwanda's P4P scheme pays facilities $2.68 for every HIV-positive woman using a modern contraceptive method. This evidence is consistent with the prediction that services that are jointly produced with highly rewarded services may increase in response to P4P; the increase in contraceptive coverage could be driven in part by such commonalities in production.

One factor that could have contributed to increases in less generously rewarded prenatal care services is the quality multiplier M: even though the explicit bonus payments for prenatal care are small, these services comprise a large share of the quality multiplier in Rwanda's program. This may mitigate multitasking between more generously rewarded health care services such as institutional deliveries and less generously rewarded prenatal care services such as tetanus vaccination and malaria prophylaxis.

While we observe no statistically significant impacts on our three prenatal measures, we can speculate about the large effect sizes for tetanus vaccinations relative to malaria prophylaxis and the antenatal care visits. Although the latter two services receive broadly Malaria prophylaxis, however, receives the same similar rewards as tetanus vaccination and is are also a component of prenatal care and hence the quality multiplier M, but we observe a smaller no significant increase improvements in these services. in this service. For malaria prophylaxis, one possible explanation for this discrepancy arises from the phrasing of the DHS questionnaire – the questionnaire asks whether women "took" any antimalarial drugs during pregnancy, not whether they were given any antimalarial drugs. There may have been an increase in the dispensation of antimalarials to pregnant women as a result of P4P, but if women chose not to take these drugs (for example due to unpleasant side effects) then this increase would not be reflected in our

results. Whereas providers have control over whether a pregnant patient receives a tetanus vaccine because they directly administer the injection, providers have far less control over whether patients will follow-through with an ongoing therapy that must be self-administered, such as antimalarials, or comply with the recommended schedule of antenatal care visits. This highlights one of the reasons why P4P initiatives, even where they are successful in improving the quality of care, often do not improve health outcomes: rewarded activities may contribute only weakly to actual health improvements, especially when providers have limited control over whether patients follow-through with therapy or recommendations. In addition to accounting for patient demand, supply-side interventions also need to consider adherence and patient behavior.

The significance of prenatal care in the quality multiplier might also explain the increases in several unrewarded prenatal care services, i.e. iron supplementation and urinalysis. Another explanation for the increase in only a few of the prenatal services could be the pre-roll-out levels of services. At baseline iron supplementation and urinalysis were provided for only 29.55% and 8.57% of pregnant women, respectively, so for both measures there was significant scope for improvement. In contrast, receipt of prenatal care by a trained provider and the likelihood of being weighed in pregnancy were both very high at baseline (i.e. 95%), so ceiling effects likely prevented any further significant improvements. Also, 74% of pregnant women reported that their blood pressure was measured at baseline – while this is low enough that ceiling effects should not hinder improvements, it is high enough that further improvements might not have increased the value of the quality multiplier and therefore might not have been worth the additional effort. At baseline, only 37.4% of pregnant women had their blood drawn and only 5.83% were told the possible complications of pregnancy, so it is unclear why these services did not significantly improve when iron supplementation and

urinalysis did. Nonetheless, these findings suggest that the quality multiplier might play a role in encouraging improvements in certain prenatal care services that are not explicitly rewarded.

The modest impacts of P4P on rewarded and unrewarded aspects of healthcare quality may explain why we fail to find any significant impacts of P4P on the health outcomes in our data. Moreover, if rewarded activities contribute only weakly to actual health improvements, then even where we observe improvements in health care quality, improvements in outcomes may not follow automatically.

The response to P4P likely varies across areas with different levels of baseline quality due to differences they face in the marginal costs of health service production, the value of the quality multiplier M, or other endogenous factors that contribute to underlying variations in the quality of care such as personnel motivation and infrastructure. These underlying factors might help explain why we find that EAs in the middle quality tier achieve improvements across a larger number of measures than EAs in the high or low tiers.

The only significant impacts of P4P that we observe in high quality tier areas are for institutional deliveries and iron supplementation in pregnancy. There are several possible explanations for why these higher-performing areas did not experience improvements in other rewarded or unrewarded services. First, there may be an increasing marginal cost of quality improvement. Facilities with higher performance at baseline might need to attract new patients, rather than changing clinical activities for existing patients, which could be associated with high marginal costs. Related, if higher quality facilities are closer to their production possibility frontier (determined by infrastructural or resource capacity), further increases in service provision could be very costly, or may require lumpy and significant

investments. Although utilization rates of basic services, such as prenatal care and deliveries, are low even in high-performing areas, these ceiling effects could be caused by shortages in human resources or basic infrastructure and facilities.

The weak response of higher quality EAs to P4P may also reflect multitasking. Higher-performing facilities have the strongest incentive to increase the output of more generously rewarded services – under Rwanda's P4P scheme, the most generously rewarded services are HIV/AIDS services, so higher-performing areas may have focused on these measures (which the DHS does not contain) at the expense of the less generously rewarded measures in our data. The weak response to P4P among the bottom quality tier EAs could be driven by the low effective level of incentives, as a result of the low values of the quality multiplier M.

Rwanda's national P4P program therefore appears to have contributed to improvements in health care quality, particularly in the areas of institutional deliveries, contraceptive coverage and prenatal care. Our findings point to several areas in which Rwanda's P4P program might be strengthened, with lessons for other provider payment schemes. Analytically, rewarding health care services that are jointly produced with other desirable services may achieve broader improvements. Rewarding only a subset of services that are jointly produced is also more cost-effective than rewarding each of these services individually, because improvements may be achieved through positive production spillovers with no need for additional financial incentives. Second, the use of a quality multiplier or scaling factor based on broader measures can potentially mitigate multitasking concerns. Third, P4P programs should consider demand-side factors.

Finally, our findings suggest that in Rwanda's case either a modified payment formula or other complementary approaches are needed to encourage higher-performing facilities to increase the output of rewarded services. In the current setting, providers that were high-performing at baseline reaped substantial rewards after P4P was rolled-out, without needing to engage in any improvement activities, thereby reducing the cost-effectiveness of the program. The appropriate policy response will depend on the reason why high-performers are stagnant. If infrastructural and human resource constraints are preventing further quality improvements, then physical expansions of the facility or investment in training additional personnel may suffice. If the marginal cost of quality improvement is increasing (e.g. if the marginal patient is more costly to recruit), then providers may respond to larger rewards. However, any effort to offer incremental performance rewards requires an understanding of the underlying production function and cost structures, which remains especially challenging in health care and even more so in developing country.

REFERENCES

Arrow, K. (1963). "Uncertainty and the Welfare Economics of Medical Care." American Economic Review **53**: 941–973.

Basinga, P., P. Gertler, et al. (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**(1421–1428).

Cameron, A., J. Gelbach, et al. (2008). "Bootstrap-Based Improvements for Inference with Clustered Errors." Review of Economics and Statistics **90**(414–427).

Campbell, S., D. Reeves, et al. (2007). "Quality of primary care in England with the introduction of pay for performance." New England Journal of Medicine **57**(2): 181-190.

Campbell, S., D. Reeves, et al. (2009). "Effects of pay for performance on the quality of primary care in England." The New England Journal of Medicine **361**: 368–378.

Christianson, J., S. Leatherman, et al. (2008). "Lessons from evaluations of purchaser pay-for-performance programs: a review of the evidence." Medical Care Research and Review **65**: 5S–35S.

Cohen, J. and P. Dupas (2010). "Free Distribution or Cost-Sharing? Evidence from a Randomized Malaria Prevention Experiment." Quarterly Journal of Economics **125**: 1–45.

Doran, T., C. Fullwood, et al. (2006). "Pay-for-performance programs in family practices in the United Kingdom. ." New England Journal of Medicine **355**(4): 375-384.

Gertler, P. and C. Vermeesch (2012). Using Performance Incentives to Improve Health Outcomes. . World Bank Working Paper Series. Washington DC.

Glazer, J., T. McGuire, et al. (2008). "Mitigating the problem of unmeasured outcomes in quality reports. ." The B.E. Journal of Economic Analysis & Policy **8**(2): 1935-1682.

Glewwe, P., N. Ilias, et al. (2010). "Teacher Incentives." American Economic Journal: Applied Economics **2**(3): 205-227.

Glickman, S., F. Ou, et al. (2007). "Pay for performance, quality of care, and outcomes in acute myocardial infarction." JAMA: Journal of the American Medical Association **297**(21): 2373-2380.

Hoff, M. (2010). Improving the incentive mechanisms in Rwanda's health care system. Cambridge, MA, Harvard University.

Holmstrom, B. and P. Milgrom (1991). "Multitask Principal–Agent Analyses: Incentive Contracts, Asset Ownership, and Job Design." Journal of Law, Economics, and Organization **7**(special issue): 24-52.

Kling, J. R., J. B. Liebman, et al. (2007). "Experimental Analysis of Neighborhood Effects." Econometrica **75**(1): 83-119.

Leonard, K. L. (2003). "African traditional healers and outcome-contingent contracts in health care." Journal of Development Economics **71**(1): 1-22.

Li, J., J. Hurley, et al. (2011). Physician response to pay-for-performance: evidence from a natural experiment. . NBER Working Paper Series. Cambridge MA.

Lindenauer, P. K., D. Remus, et al. (2007). "Public Reporting and Pay for Performance in Hospital Quality Improvement." New England Journal of Medicine **356**(5): 486-496.

Lu, C., B. Chin, et al. (2012). "Towards Universal Health Coverage: An Evaluation of Rwanda Mutuelles in Its First Eight Years." PLoS ONE **7**(6): e39282.

McGuire, T. G. and M. V. Pauly (1991). "Physician response to fee changes with multiple payers." Journal of Health Economics **10**(4): 385-410.

Ministry of Health of Rwanda (2010). Rwanda Community-Based Health Insurance Policy. Kigali, Rwanda.

Mullen, K. J., R. G. Frank, et al. (2010). "Can you get what you pay for? Pay-for-performance and the quality of healthcare providers." The RAND Journal of Economics **41**(1): 64-91.

Muralidharan, K. and V. Sundararaman (2011). "Teacher Performance Pay: Experimental Evidence from India." Journal of Political Economy **119**(1): 39-77.

National Institute of Statistics of Rwanda (NISR) and Macro International Inc (2005). Rwanda Demographic and Health Survey I. Macro International. Calverton, United States: .

Olken, B., J. Onishi, et al. (2011). Should Aid Reward Performance? Evidence from a field experiment on health and education in Indonesia. NBER Working Paper Cambridge MA.

Peabody, J., R. Shimkhada, et al. (2011). "Financial Incentives And Measurement Improved Physicians' Quality Of Care In The Philippines." Health Affairs **30**(4): 773-781.

Petersen, L. A., L. D. Woodard, et al. (2006). "Does Pay-for-Performance Improve the Quality of Health Care?" Annals of Internal Medicine **145**(4): 265-272.

Prendergast, C. (1999). "The Provision of Incentives in Firms." Journal of Economic Literature **37**(1): 7-63.

Roos, N. M., M. Wiegman, et al. (2012). "Visual disturbances in (pre)eclampsia." Obstetrical and Gynecological Survey **67**(4): 242–250.

Rosenthal, M. B. and A. Dudley (2007). "Pay-for-Performance: Will the Latest Payment Trend Improve Care?" JAMA: Journal of the American Medical Association **297**.

Rosenthal, M. B. and R. G. Frank (2006). "What is the empirical basis for paying for quality in health care?" Medical Care Research and Review **63**: 135–157.

Rosenthal, M. B., R. G. Frank, et al. (2005). "Early experience with pay-for-performance: from concept to practice." JAMA: Journal of the American Medical Association **294**: 1788–1793.

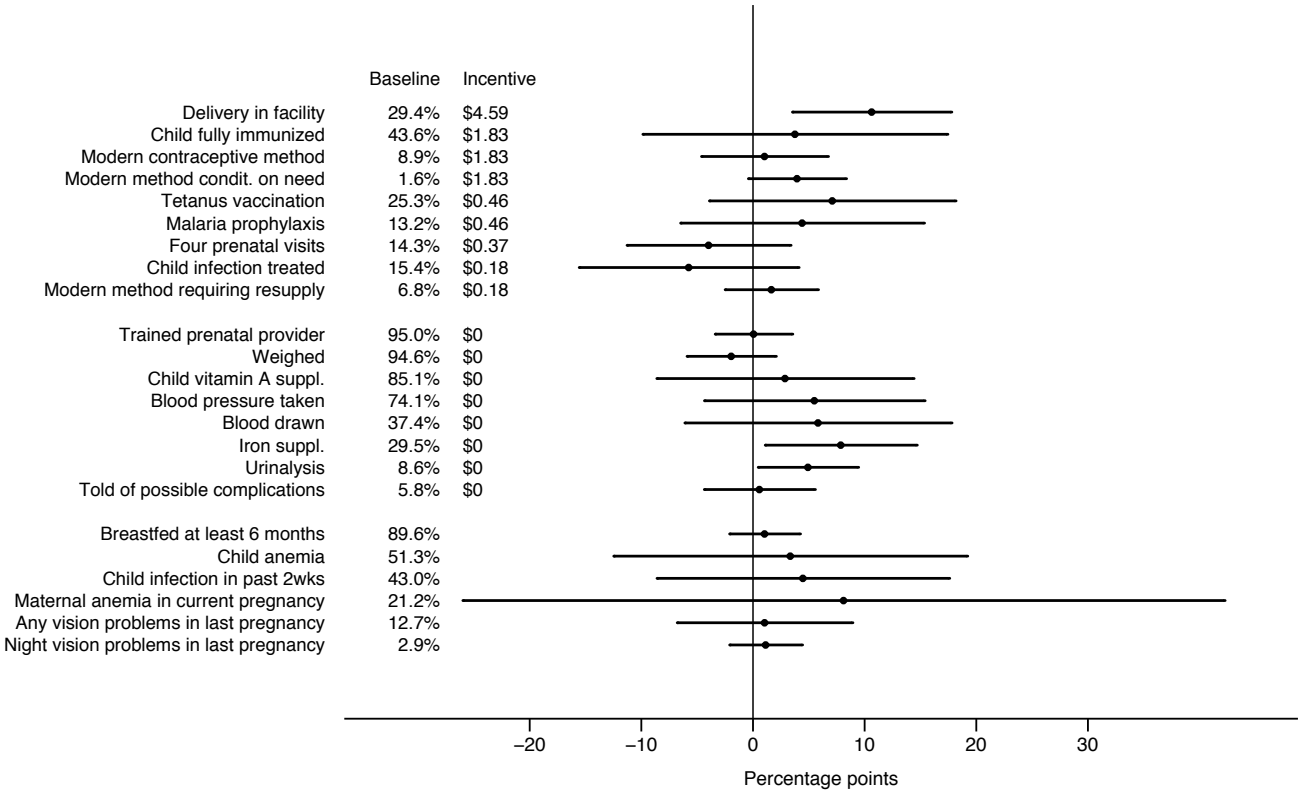
Sherry, T. B. (2012). A note on the comparative statics of pay-for-performance in health care. .

Steel, N., S. Maisey, et al. (2007). "Quality of clinical primary care and targeted incentive payments: an observational study." British Journal of General Practice **57**: 449–454.

Sutton, M., R. Elder, et al. (2010). "Record rewards: The effects of targeted quality incentives on the recording of risk factors by primary care providers." Health Economics **19**: 1-13.

Werner, R., J. T. Kolstad, et al. (2011). "The effect of pay-for-performance in hospitals: lessons for quality improvement. ." Health Affairs **30**(4): 690-698. .
World Health Organization (2012). WHO National Health Accounts Database. Geneva, World Health Organization.

Figure 1: Estimated treatment effects for rewarded services, unrewarded services and outcomes



Coefficients for post*treat with 95% CIs based on a t(G-2) distribution. Positive estimates represent improvement for services in the top and middle panel. Vision problems can be a symptom of pre-eclampsis/gestational diabetes. Night blindness is a specific symptom of vitamin A deficiency.

Table 1: Summary statistics for pre-intervention period

Outcomes (binary)	N	Mean Control	Mean Treat	p-value
Rewarded services				
Delivery in facility	4,914	0.30	0.29	0.64
Tetanus vaccination	2,723	0.27	0.24	0.22
Malaria prophylaxis	2,749	0.13	0.13	0.95
Four prenatal visits	2,748	0.13	0.15	0.42
Child fully immunized	3,988	0.47	0.41	0.40
Child infection treated	1,108	0.13	0.17	0.19
Modern contraceptive method	1,814	0.10	0.08	0.33
Modern method condit. on need	684	0.01	0.02	0.70
Modern method requiring resupply	1,814	0.08	0.06	0.12
Unrewarded services				
Trained prenatal provider	2,755	0.96	0.94	0.19
Weighed (prenatal care)	2,628	0.94	0.95	0.47
Blood pressure taken (prenatal care)	2,627	0.76	0.73	0.41
Blood drawn (prenatal care)	2,626	0.40	0.36	0.49
Iron suppl. (prenatal care)	2,728	0.34	0.27	0.22
Urinalysis (prenatal care)	2,613	0.08	0.09	0.49
Told of possible complic. (prenatal care)	2,624	0.06	0.06	0.98
Child vitamin A suppl.	1,965	0.85	0.85	1.00
Outcomes				
Breastfed at least 6 months	4,496	0.90	0.89	0.79
Any vision problems in last pregnancy	2,739	0.14	0.12	0.56
Night vision problems in last pregnancy	2,739	0.04	0.02	0.13
Maternal anemia in current pregnancy	198	0.20	0.22	0.83
Child anemia	1,268	0.53	0.50	0.64
Child infection in past 2wks	2,577	0.43	0.43	0.95
Covariates				
Mother's age	4,914	31.23	31.20	0.91
Mother's education	4,914	3.57	3.46	0.65
Hhold size	4,914	5.69	5.62	0.52
N children $\leq$ 5 years	4,914	1.87	1.85	0.70
N births	4,914	4.32	4.47	0.30
Wealth Quintile	4,914	1.91	1.73	0.30
Married/cohabitating	4,914	0.88	0.87	0.52
Insured	4,914	0.53	0.57	0.56
EA baseline quality Q: 1(low) to 3(high)	153	2.00	1.99	0.94

See Appendix for details on sample definitions and variable constructions. Vision problems are symptomatic of pre-eclampsis/gestational diabetes. Night blindness is a specific symptom of vitamin A deficiency. Covariate values based on estimation sample for facility deliveries. p-values clustered on district level based on t-distribution with G-2 d.f.

Table 2: Effect of P4P on rewarded services (percentage point changes)

		Prenatal Care			Child health		Modern contraception		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	Delivery in facility	Tetanus vaccination	Malaria prophylaxis	Four prenatal visits	Child fully immunized	Child infection treated	Use of modern method	Modern method condit. on need	Method requiring resupply
Incentive in USD	4.59	0.46	0.46	0.37	1.83	0.18	1.83	1.83	0.18
Pre-period mean (%)	29.43	25.27	13.24	14.30	43.63	15.43	8.88	1.61	6.84
A. Average effect (results for covariates omitted)									
Treat*Post	10.63***	7.10	4.40	-3.98	3.75	-5.76	1.03	3.94*	1.64
	(3.36)	(5.22)	(5.16)	(3.46)	(6.45)	(4.65)	(2.68)	(2.06)	(1.96)
Post	-1.52	-1.16	-1.32	5.63	8.50	-1.23	16.27***	1.83	14.72***
	(3.84)	(4.54)	(4.73)	(4.00)	(5.45)	(6.51)	(1.98)	(1.33)	(1.42)
B. Effect by baseline quality Q									
Treat*Post Q-low	8.12	0.95	1.05	-7.67*	-2.09	-7.17	0.87	8.85**	-2.45
	(5.10)	(5.17)	(8.85)	(4.12)	(8.32)	(6.38)	(4.30)	(3.56)	(4.39)
Treat*Post Q-medium	7.54	12.05	11.58**	-6.28	4.30	-0.08	5.90*	5.18*	10.16***
	(4.45)	(9.31)	(5.47)	(6.22)	(7.36)	(6.55)	(3.32)	(2.89)	(2.89)
Treat*Post Q-high	13.30**	9.36	-2.00	-0.68	5.73	-10.85	-2.79	-0.97	-1.66
	(5.56)	(6.47)	(7.22)	(5.05)	(11.36)	(8.76)	(3.80)	(3.69)	(3.56)
Post Q-low	-1.65	7.48	-3.37	5.12	10.74*	2.26	17.60***	-1.32	17.77***
	(4.49)	(4.61)	(8.05)	(4.77)	(6.05)	(7.25)	(2.30)	(2.21)	(2.74)
Post Q-medium	5.44	-5.13	-1.82	11.83**	14.77**	-3.86	10.67***	-0.37	7.36***
	(4.92)	(8.46)	(3.93)	(5.58)	(5.46)	(5.85)	(2.54)	(0.66)	(2.22)
Post Q-high	-6.53	-6.82	1.56	1.87	1.42	-1.84	19.56***	6.49***	17.80***
	(4.63)	(4.74)	(5.46)	(4.33)	(10.17)	(9.96)	(2.32)	(2.13)	(2.66)
Mother's age	-0.10	-0.58***	0.08	0.35*	0.31	-0.11	-0.55***	-0.05	-0.58***
	(0.15)	(0.18)	(0.16)	(0.17)	(0.23)	(0.20)	(0.13)	(0.11)	(0.13)
Mother's education	1.55***	-0.27	0.69**	0.08	0.42	0.46	1.20***	0.61***	0.69***
	(0.18)	(0.22)	(0.29)	(0.28)	(0.30)	(0.50)	(0.18)	(0.21)	(0.20)
Wealth Quintile	2.56***	1.19*	0.19	1.02*	0.75	1.40	1.79***	0.33	1.57***
	(0.54)	(0.65)	(0.68)	(0.52)	(0.51)	(0.85)	(0.56)	(0.49)	(0.49)
Insured	7.77***	2.22	1.73	2.66**	3.69**	10.43***	1.77	1.34	1.39
	(1.17)	(1.55)	(1.63)	(1.26)	(1.41)	(1.46)	(1.30)	(1.23)	(1.08)
Additional covars	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Birth year FE	Yes	Yes	Yes	Yes	Yes	Yes	No	No	No
Area FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N†	6,144	3,945	3,978	3,983	4,588	2,276	3,709	1,173	3,709
R2 (quartile models)	0.17	0.10	0.26	0.04	0.08	0.04	0.11	0.06	0.09

* p<0.10, ** p<0.05, *** p<0.01 OLS models, standard errors clustered within districts. Scaled by 100 for exposition. Panel B uses same models as panel A, replacing treat and post*treat with quartile interactions. † Details on varying reference population and covariate sets in Data section and Appendix. Samples in cols 7-9 restricted to married/cohabitating women. Models on child health also control for child's age and gender. Child infection treated at government facility.

Table 3: Effect of P4P on unrewarded services (percentage point changes)

	Prenatal Care							
	(1) Trained prenatal provider	(2) Weighed	(3) Blood pressure taken	(4) Blood drawn	(5) Iron suppl.	(6) Urinalysis	(7) Told of possible complica-tions	(8) Child vitamin A suppl.
Incentive in USD	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Pre-period mean (%)	95.03	94.60	74.08	37.40	29.55	8.57	5.83	85.14
A. Average effect (results for covariates omitted)								
Treat*Post	0.05	-1.95	5.49	5.82	7.86**	4.92**	0.57	2.87
	(1.63)	(1.88)	(4.67)	(5.65)	(3.21)	(2.11)	(2.34)	(5.44)
Post	-1.67	5.05**	1.37	-3.00	-0.27	-1.82	0.47	-2.34
	(1.40)	(1.87)	(4.35)	(4.62)	(4.29)	(2.82)	(3.02)	(5.84)
B. Effect by baseline quality Q								
Treat*Post Q-low	-2.41	-3.78	1.79	0.65	-0.14	3.57	3.29	-2.12
	(2.24)	(3.88)	(8.03)	(9.79)	(6.39)	(3.31)	(3.77)	(9.11)
Treat*Post Q-medium	0.82	0.13	13.09**	8.46	12.63	6.97**	3.53	10.14
	(2.13)	(1.89)	(5.56)	(8.20)	(8.04)	(3.14)	(3.99)	(7.15)
Treat*Post Q-high	1.38	-1.27	0.26	9.59	12.23*	4.48	-5.78*	0.37
	(1.90)	(2.24)	(5.23)	(6.58)	(6.87)	(4.59)	(2.81)	(7.95)
Post Q-low	-0.89	8.72***	9.93*	3.32	7.57	-2.70	1.33	-1.69
	(1.57)	(2.59)	(5.69)	(7.50)	(6.51)	(2.86)	(4.22)	(8.07)
Post Q-medium	-1.18	2.72	-0.75	-5.66	-4.00	-3.45	-0.68	-4.80
	(1.64)	(2.01)	(5.40)	(6.91)	(6.59)	(2.92)	(3.68)	(6.88)
Post Q-high	-2.74	3.02	-5.36	-7.47	-5.24	0.27	0.56	-0.53
	(1.60)	(1.94)	(3.40)	(5.57)	(5.90)	(3.71)	(2.60)	(7.39)
Mother's age	-0.16*	0.20*	0.46**	0.21	0.35	0.21	0.01	0.01
	(0.09)	(0.10)	(0.19)	(0.18)	(0.21)	(0.14)	(0.08)	(0.15)
Mother's education	0.19	0.26***	0.85***	-0.21	0.41**	0.47*	0.29**	0.10
	(0.17)	(0.08)	(0.25)	(0.30)	(0.19)	(0.24)	(0.13)	(0.26)
Wealth Quintile	0.46	0.29	1.89**	0.90	1.86***	0.76	-0.04	0.77
	(0.30)	(0.34)	(0.75)	(0.91)	(0.58)	(0.52)	(0.24)	(0.68)
Insured	1.24*	0.39	1.61	2.80*	1.41	0.67	-0.65	3.38**
	(0.67)	(0.63)	(1.58)	(1.42)	(1.62)	(1.11)	(0.95)	(1.57)
Additional covars	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Birth year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Area FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N†	3,995	3,830	3,829	3,827	3,959	3,803	3,823	3,856
R2 (quartile models)	0.02	0.02	0.04	0.21	0.02	0.04	0.01	0.01

* p<0.10, ** p<0.05, *** p<0.01 OLS models, standard errors clustered within districts. Scaled by 100 for exposition. Panel B uses same models as panel A, replacing treat and post*treat with quartile interactions. † Details on varying reference population and covariate sets in Data section and Appendix. Models on child's vitamin A also control for child's age and gender.

Table 4: Effect of P4P on outcomes (percentage point changes)

		Adverse events				
	(1)	(2)	(3)	(4)	(5)	(6)
	Breastfed at least 6 months	Any vision problems in last pregnancy	Night vision problems in last pregnancy	Maternal anemia in current pregnancy	Child anemia	Child infection in past 2wks
Incentive in USD						
Pre-period mean (%)	89.59	12.67	2.92	21.21	51.26	43.00
A. Average effect (results for covariates omitted)						
Treat*Post	1.04	1.04	1.13	8.11	3.34	4.47
	(1.49)	(3.71)	(1.53)	(16.16)	(7.49)	(6.19)
Post	1.72	-1.87	-1.95	-7.32	-28.55***	-17.77***
	(1.96)	(2.72)	(1.71)	(13.66)	(5.46)	(5.40)
B. Effect by baseline quality Q						
Treat*Post Q-low	-0.21	-1.71	1.26	25.75	8.30	10.56**
	(2.38)	(4.69)	(2.18)	(15.04)	(12.71)	(4.02)
Treat*Post Q-medium	4.95*	5.50	0.41	-18.78	3.23	17.01*
	(2.72)	(4.07)	(2.36)	(27.68)	(8.03)	(8.25)
Treat*Post Q-high	-2.22	1.00	2.97	-4.42	-1.87	-14.74
	(4.23)	(4.20)	(2.45)	(10.35)	(9.01)	(11.36)
Post Q-low	1.93	3.21	-0.73	-18.48	-39.02***	-21.72***
	(2.19)	(2.42)	(1.60)	(13.89)	(8.09)	(2.09)
Post Q-medium	0.37	-6.97**	-3.30	14.19	-24.14***	-25.30***
	(3.21)	(3.26)	(2.49)	(25.33)	(7.39)	(7.63)
Post Q-high	2.68	-3.23	-2.31	2.49	-23.07***	-7.48
	(3.06)	(3.21)	(2.26)	(5.57)	(6.61)	(9.10)
Mother's age	0.06	0.13	0.07	1.64***	-0.44**	0.04
	(0.09)	(0.14)	(0.08)	(0.54)	(0.21)	(0.18)
Mother's education	0.02	0.11	-0.10	0.23	-0.02	-0.36
	(0.16)	(0.26)	(0.12)	(1.27)	(0.21)	(0.24)
Wealth Quintile	0.05	-0.03	-0.13	-0.01	-0.90*	-0.60
	(0.22)	(0.36)	(0.18)	(2.13)	(0.45)	(0.91)
Insured	-1.56	0.15	0.88	1.35	-5.64**	0.89
	(1.01)	(1.52)	(0.71)	(4.31)	(1.95)	(1.65)
Additional covars	Yes	Yes	Yes	Yes	Yes	Yes
Birth year FE	Yes	Yes	Yes	Yes	Yes	Yes
Area FE	Yes	Yes	Yes	Yes	Yes	Yes
N†	5,430	3,972	3,972	492	3,734	5,175
R2 (quartile models)	0.06	0.01	0.01	0.03	0.07	0.03

* p<0.10, ** p<0.05, *** p<0.01 OLS models, standard errors clustered within districts. Scaled by 100 for exposition. Panel B uses same models as panel A, replacing treat and post*treat with quartile interactions. † Details on varying reference population and covariate sets in Data section and Appendix. Models on child's anemia and infection also control for child's age and gender.

A Paper Appendix

A.1 Treatment assignment

In Phase 1 of the national P4P experimental roll-out, between June and October 2006, the 19 districts that had not previously participated in a P4P pilot were organized into 8 groups. Districts in each group had similar rainfall and population density, and their inhabitants had similar sources of livelihoods. Half the districts in each group were randomly assigned to treatment and the remaining half to control status, resulting in 10 treatment and 9 control districts. Prior to the roll-out of P4P, the government of Rwanda implemented a decentralization initiative that involved redrawing district boundaries. Consequently, 2 control districts were combined with Phase 0 districts that had existing P4P pilots and were thus reassigned to treatment status, resulting in 12 treatment districts and 7 control districts in Phase 1. Further details are described in Basinga et al. (2011).

A.2 Combining the DHS waves

Since the two DHS waves are collected in the same EAs, the respondents may overlap in the two surveys, so that births before the earlier wave may be double-counted. We identified households within an EA that are potentially interviewed in both waves by using the respondent's birthdate, and her children's birthdates, sex and twin status (for children born before January 2005, before the first wave). We conservatively assume that all respondents who cannot be uniquely identified across waves may be duplicates. We delete from the 2005 estimation sample the less than 1 percent of mothers and children who may have been re-interviewed in the 2007 wave.

Table A.1: Rewarded indicators and components of quality multiplier M

Rewarded output and quality indicators	Payment/unit (US $)	Available in DHS
Primary Care		
Emergency referrals during curative treatment	1.83	
Curative care visits	0.18	Y
Family Planning		
First-time family planning visits	1.83	Y[†]
1-month contraceptive resupply	0.18	Y[†]
Maternal Health		
Deliveries in the facility	4.59	Y
Emergency transfers to hospital for obstetric care during delivery	4.59	
At-risk pregnancies referred to hospital for delivery during prenatal care	1.83	
Women who received 2nd dose of malaria prophylaxis during prenatal care	0.46	Y
Women who received appropriate tetanus vaccination during prenatal care	0.46	Y
Women who completed 4 prenatal care visits	0.37	Y
First prenatal care visits	0.09	
Child Health		
Malnourished children referred for treatment during preventive care visit	1.83	
Children who completed vaccination on time	1.83	Y
Child (0-59 months) preventive care visits	0.18	
HIV/AIDS		
PMTCT: exposed children tested	8.93	
New pediatric clients put on ARVs	6.70	
Prevention of mother-to-child-transmission (PMTCT): partner tested	4.58	
PMTCT: women under treatment with ARVs during labor	4.58	
New adult clients put on ARVs	4.58	
HIV+ clients tested for CD4 count	4.58	
HIV+ women who use modern method of family planning	2.68	
HIV+ clients tested for TB	2.68	
Voluntary counseling and testing	0.89	
HIV+ clients treated with cotrimoxazole each month	0.44	

Components of the quality multiplier M	Weight in multiplier	Share of structural factors (rather than process factors)
Curative Care	0.170	0.23
Delivery	0.130	0.40
Prenatal care	0.126	0.12
Family Planning	0.114	0.22
HIV Services	0.090	1.00
Immunization	0.070	0.40
Pharmacy Management	0.060	1.00
General Administration	0.052	1.00
Financial Management	0.050	1.00
Preventive Care	0.052	0.15
Laboratory Services	0.030	1.00
Cleanliness	0.028	1.00
TB Services	0.028	0.28

[†] First-time family planning visit includes prescription of modern contraceptive method, medical history and physical exam; DHS records "current use of modern method" which is used as proxy here. Contraceptive resupply measured as current use of a modern method requiring regular resupplies.

Table A.2: Sample and variable definitions for outcome measures

Outcome	Sample and variable definitions
Rewarded services	
Delivery in facility	All births in the 5 years preceding each survey wave. 1 if birth occurred in a public or private health facility; 0 if not.
Tetanus vaccination	Most recent birth in the 5 years preceding each survey wave. 1 if at least two tetanus vaccine injections were given during the pregnancy; 0 if fewer than 2 injections were given.
Malaria prophylaxis	Most recent birth in the 5 years preceding each survey wave. 1 if any antimalarial drugs were taken during the pregnancy; 0 if not.
Four prenatal care visits	Most recent birth in the 5 years preceding each survey wave. 1 if 4 or more prenatal care visits were completed during the pregnancy; 0 if fewer than 4 visits.
Child fully immunized	Children age 12-59 months. 1 if fully immunized by age 1 (i.e. received BCG, measles vaccine, 4 doses of oral polio vaccine & 3 doses of DPT vaccine); 0 if not.
Child infection treated	All study children who had an infection (i.e. diarrheal disease, fever or cough) in the 2 weeks preceding the survey. 1 if treated for infection at a health facility; 0 if not.
Modern contraceptive method	All women married or cohabitating with a partner. 1 if currently using modern contraception; 0 if not.
Modern method condit. on need	All fecund women married or cohabitating with a partner, who are not currently pregnant, breastfeeding or amenorrheic, and who report that they wish to limit or space births. 1 if currently using modern contraception; 0 if not.
Modern method req. resupply	All women married or cohabitating with a partner. 1 if currently using a modern contraceptive method requiring regular re-supply (i.e. pill, injectable, condom or female condom, spermicide); 0 if not.
Unrewarded services	
Prenatal care (7 measures)	Most recent birth in the 5 years preceding each survey wave. 7 prenatal measures: weighed in pregnancy, blood pressure measured, urinalysis performed, blood drawn, given iron supplement, told of possible complications in pregnancy. Each individual measure equals 1 if the service/procedure was received; 0 if not.
Child vitamin A suppl.	Children age 12-59 months. 1 if received vitamin A supplement in the past 6 months; 0 if no supplement.
Health outcomes	
Breastfed at least 6 months	All children who are either alive and older than 6 months, or dead but died after 6 months of age. 1 if child was breastfed for at least 6 months.
Any vision problems	Daytime vision difficulties can be symptomatic of advanced gestational diabetes or pre-eclampsia; nighttime vision difficulties result from vitamin A deficiency. Most recent birth in the 5 years preceding each survey wave. 1 if experienced difficulties with daytime or nighttime vision during pregnancy; 0 if no vision difficulties.
Night vision problems	Nighttime vision difficulties result from vitamin A deficiency. Most recent birth in the 5 years preceding each survey wave. 1 if experienced difficulties with nighttime vision only during pregnancy; 0 if experienced daytime vision difficulties or no vision difficulties.
Maternal anemia	Random sample of 50% of pregnant women surveyed in DHS III; all pregnant women surveyed in DHS IV. 1 if classified as mild, moderate or severe anemia; 0 if classified as no anemia. See DHS documentation for clinical protocols.
Child infection	All study children who had an infection (i.e. diarrheal disease, fever or cough) in the 2 weeks preceding the survey. 1 if treated for infection at a health facility; 0 if not.
Child anemia	Random sample of 50% of the children surveyed in DHS III; all children surveyed in DHS IV. 1 if classified as mild, moderate or severe anemia; 0 if classified as no anemia. See DHS documentation for clinical protocols.

B Online Appendix

Table B.1: Effect of P4P on rewarded services (percentage point changes)

	Prenatal Care				Child health		Modern contraception		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	Delivery in facility	Tetanus vaccination	Malaria prophylaxis	Four prenatal visits	Child fully immunized	Child infection treated	Use of modern method	Modern method condit. on need	Method requiring resupply
A. Average effect									
Treat*Post	10.63***	7.10	4.40	-3.98	3.75	-5.76	1.03	3.94*	1.64
	(3.36)	(5.22)	(5.16)	(3.46)	(6.45)	(4.65)	(2.68)	(2.06)	(1.96)
Post	-1.52	-1.16	-1.32	5.63	8.50	-1.23	16.27***	1.83	14.72***
	(3.84)	(4.54)	(4.73)	(4.00)	(5.45)	(6.51)	(1.98)	(1.33)	(1.42)
Mother's age	-0.10	-0.59***	0.09	0.35*	0.30	-0.12	-0.55***	-0.05	-0.58***
	(0.15)	(0.18)	(0.16)	(0.17)	(0.23)	(0.20)	(0.13)	(0.10)	(0.13)
Mother's education	1.55***	-0.27	0.69**	0.08	0.43	0.45	1.20***	0.62**	0.70***
	(0.18)	(0.22)	(0.28)	(0.28)	(0.30)	(0.49)	(0.18)	(0.21)	(0.20)
Hhold size	1.04**	-0.97*	0.02	0.49	0.13	-0.46	1.47***	-0.65*	1.06**
	(0.47)	(0.53)	(0.37)	(0.48)	(0.48)	(0.79)	(0.49)	(0.31)	(0.38)
N children $\leq$ 5 years	-3.44***	-2.42**	-2.02	-1.47	0.22	-0.22	1.92**	2.85***	1.64**
	(1.17)	(0.88)	(1.37)	(0.99)	(1.31)	(1.35)	(0.67)	(0.91)	(0.68)
N births	-3.01***	-3.52***	-0.13	-1.58**	-1.13*	0.74	1.21***	0.45	1.35***
	(0.58)	(0.67)	(0.52)	(0.57)	(0.60)	(0.94)	(0.40)	(0.38)	(0.42)
Wealth Quintile	2.51***	1.23*	0.09	0.96*	0.73	1.37	1.81***	0.36	1.58***
	(0.53)	(0.65)	(0.66)	(0.52)	(0.51)	(0.86)	(0.57)	(0.48)	(0.50)
Married/cohabitating	4.40**	-2.77	0.01	3.76**	8.20***	-2.36			
	(1.84)	(1.91)	(2.32)	(1.63)	(2.47)	(1.93)			
Insured	7.84***	2.28	1.87	2.70**	3.74**	10.51***	1.66	1.32	1.37
	(1.17)	(1.53)	(1.69)	(1.26)	(1.40)	(1.49)	(1.28)	(1.25)	(1.03)
Child's age					-9.01***	-1.94			
					(0.72)	(1.97)			
Male child					0.26	0.34			
					(1.18)	(1.62)			
o.Married/cohabitating							0.00	0.00	0.00
							(.)	(.)	(.)
Birth year FE	Yes	Yes	Yes	Yes	Yes	Yes	No	No	No
District FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N†	6,144	3,945	3,978	3,983	4,588	2,276	3,709	1,173	3,709
R2	0.17	0.11	0.26	0.03	0.08	0.04	0.10	0.05	0.08
Pre-period mean (%)	29.43	25.27	13.24	14.30	43.63	15.43	8.88	1.61	6.84
Incentive in USD	4.59	0.46	0.46	0.37	1.83	0.18	1.83	1.83	0.18

* p<0.10, ** p<0.05, *** p<0.01 OLS models, standard errors clustered within districts. Scaled by 100 for exposition. † Details on varying reference population and covariate sets in Data section and Appendix. Samples in cols 7-9 restricted to married/cohabitating women. Models on child health also control for child's age and gender. Child infection treated at government facility.

Table B.2: Effect of P4P on rewarded services (percentage point changes)

	Prenatal Care				Child health		Modern contraception		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	Delivery in facility	Tetanus vaccination	Malaria prophylaxis	Four prenatal visits	Child fully immunized	Child infection treated	Use of modern method	Modern method condit. on need	Method requiring resupply
B. Effect by baseline quality Q (results for covariates omitted)									
Treat*Post Q-low	8.12	0.95	1.05	-7.67*	-2.09	-7.17	0.87	8.85**	-2.45
	(5.10)	(5.17)	(8.85)	(4.12)	(8.32)	(6.38)	(4.30)	(3.56)	(4.39)
Treat*Post Q-medium	7.54	12.05	11.58**	-6.28	4.30	-0.08	5.90*	5.18*	10.16***
	(4.45)	(9.31)	(5.47)	(6.22)	(7.36)	(6.55)	(3.32)	(2.89)	(2.89)
Treat*Post Q-high	13.30**	9.36	-2.00	-0.68	5.73	-10.85	-2.79	-0.97	-1.66
	(5.56)	(6.47)	(7.22)	(5.05)	(11.36)	(8.76)	(3.80)	(3.69)	(3.56)
Post Q-low	-1.65	7.48	-3.37	5.12	10.74*	2.26	17.60***	-1.32	17.77***
	(4.49)	(4.61)	(8.05)	(4.77)	(6.05)	(7.25)	(2.30)	(2.21)	(2.74)
Post Q-medium	5.44	-5.13	-1.82	11.83**	14.77**	-3.86	10.67***	-0.37	7.36***
	(4.92)	(8.46)	(3.93)	(5.58)	(5.46)	(5.85)	(2.54)	(0.66)	(2.22)
Post Q-high	-6.53	-6.82	1.56	1.87	1.42	-1.84	19.56***	6.49***	17.80***
	(4.63)	(4.74)	(5.46)	(4.33)	(10.17)	(9.96)	(2.32)	(2.13)	(2.66)
Mother's age	-0.10	-0.58***	0.08	0.35*	0.31	-0.11	-0.55***	-0.05	-0.58***
	(0.15)	(0.18)	(0.16)	(0.17)	(0.23)	(0.20)	(0.13)	(0.11)	(0.13)
Mother's education	1.55***	-0.27	0.69**	0.08	0.42	0.46	1.20***	0.61***	0.69***
	(0.18)	(0.22)	(0.29)	(0.28)	(0.30)	(0.50)	(0.18)	(0.21)	(0.20)
Hhold size	1.03**	-0.99*	0.02	0.49	0.12	-0.45	1.46***	-0.66**	1.06**
	(0.46)	(0.54)	(0.38)	(0.48)	(0.49)	(0.79)	(0.49)	(0.31)	(0.38)
N children $\leq$ 5 years	-3.44***	-2.35**	-1.98	-1.46	0.23	-0.30	1.92**	2.84***	1.65**
	(1.17)	(0.88)	(1.36)	(0.98)	(1.31)	(1.33)	(0.67)	(0.91)	(0.68)
N births	-3.01***	-3.54***	-0.09	-1.59**	-1.12*	0.72	1.23***	0.44	1.37***
	(0.59)	(0.66)	(0.52)	(0.58)	(0.59)	(0.93)	(0.40)	(0.39)	(0.41)
Wealth Quintile	2.56***	1.19*	0.19	1.02*	0.75	1.40	1.79***	0.33	1.57***
	(0.54)	(0.65)	(0.68)	(0.52)	(0.51)	(0.85)	(0.56)	(0.49)	(0.49)
Married/cohabitating	4.51**	-2.94	0.17	3.93**	8.19***	-2.37			
	(1.82)	(1.94)	(2.33)	(1.63)	(2.44)	(2.01)			
Insured	7.77***	2.22	1.73	2.66**	3.69**	10.43***	1.77	1.34	1.39
	(1.17)	(1.55)	(1.63)	(1.26)	(1.41)	(1.46)	(1.30)	(1.23)	(1.08)
Child's age					-9.04***	-2.00			
					(0.72)	(2.05)			
Male child					0.19	0.35			
					(1.18)	(1.62)			
o.Married/cohabitating							0.00	0.00	0.00
							(.)	(.)	(.)
Birth year FE	Yes	Yes	Yes	Yes	Yes	Yes	No	No	No
District FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N†	6,144	3,945	3,978	3,983	4,588	2,276	3,709	1,173	3,709
R2	0.17	0.10	0.26	0.04	0.08	0.04	0.11	0.06	0.09

* p<0.10, ** p<0.05, *** p<0.01 OLS models, standard errors clustered within districts. Scaled by 100 for exposition. † Details on varying reference population and covariate sets in Data section and Appendix. Samples in cols 7-9 restricted to married/cohabitating women. Models on child health also control for child's age and gender. Child infection treated at government facility.

Table B.3: Effect of P4P on unrewarded services (percentage point changes)

	Prenatal Care							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Trained prenatal provider	Weighed	Blood pressure taken	Blood drawn	Iron suppl.	Urinalysis	Told of possible complica-tions	Child vitamin A suppl.
A. Average effect								
Treat*Post	0.05	-1.95	5.49	5.82	7.86**	4.92**	0.57	2.87
	(1.63)	(1.88)	(4.67)	(5.65)	(3.21)	(2.11)	(2.34)	(5.44)
Post	-1.67	5.05**	1.37	-3.00	-0.27	-1.82	0.47	-2.34
	(1.40)	(1.87)	(4.35)	(4.62)	(4.29)	(2.82)	(3.02)	(5.84)
Mother's age	-0.16*	0.20*	0.45**	0.21	0.35	0.21	0.01	0.01
	(0.09)	(0.10)	(0.19)	(0.18)	(0.21)	(0.14)	(0.08)	(0.15)
Mother's education	0.19	0.26***	0.86***	-0.22	0.41**	0.47*	0.30**	0.09
	(0.17)	(0.08)	(0.25)	(0.29)	(0.19)	(0.24)	(0.14)	(0.26)
Hhold size	0.21	0.17	0.34	0.57	-0.59	0.64	-0.08	-0.13
	(0.25)	(0.21)	(0.47)	(0.46)	(0.67)	(0.47)	(0.22)	(0.68)
N children $\leq$ 5 years	-0.27	-1.49**	-2.11*	-3.53***	-2.10	-3.19***	-0.77	0.41
	(0.42)	(0.56)	(1.12)	(0.97)	(1.30)	(0.64)	(0.72)	(1.13)
N births	-0.08	-0.37	-0.59	-0.82	0.10	-0.63	0.73**	-0.02
	(0.27)	(0.30)	(0.53)	(0.61)	(0.46)	(0.45)	(0.31)	(0.43)
Wealth Quintile	0.45	0.31	1.84**	0.94	1.90***	0.77	-0.07	0.72
	(0.30)	(0.34)	(0.78)	(0.91)	(0.61)	(0.53)	(0.25)	(0.69)
Married/cohabitating	3.69***	-0.33	0.72	-1.43	-0.87	-2.14	0.53	-0.13
	(1.26)	(0.87)	(2.27)	(2.16)	(2.16)	(1.72)	(1.08)	(1.87)
Insured	1.25*	0.41	1.85	2.83*	1.44	0.65	-0.55	3.50**
	(0.68)	(0.64)	(1.64)	(1.41)	(1.68)	(1.13)	(0.93)	(1.63)
Child's age								-0.40
								(1.41)
Male child								2.37**
								(1.07)
Birth year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
District FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N†	3,995	3,830	3,829	3,827	3,959	3,803	3,823	3,856
R2	0.02	0.02	0.05	0.21	0.03	0.04	0.01	0.01
Pre-period mean (%)	95.03	94.60	74.08	37.40	29.55	8.57	5.83	85.14
Incentive in USD	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

* p<0.10, ** p<0.05, *** p<0.01 OLS models, standard errors clustered within districts. Scaled by 100 for exposition. † Details on varying reference population and covariate sets in Data section and Appendix. Models on child's vitamin A also control for child's age and gender.

Table B.4: Effect of P4P on unrewarded services (percentage point changes)

	Prenatal Care							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Trained prenatal provider	Weighed	Blood pressure taken	Blood drawn	Iron suppl.	Urinalysis	Told of possible complications	Child vitamin A suppl.
B. Effect by baseline quality Q (results for covariates omitted)								
Treat*Post Q-low	-2.41	-3.78	1.79	0.65	-0.14	3.57	3.29	-2.12
	(2.24)	(3.88)	(8.03)	(9.79)	(6.39)	(3.31)	(3.77)	(9.11)
Treat*Post Q-medium	0.82	0.13	13.09**	8.46	12.63	6.97**	3.53	10.14
	(2.13)	(1.89)	(5.56)	(8.20)	(8.04)	(3.14)	(3.99)	(7.15)
Treat*Post Q-high	1.38	-1.27	0.26	9.59	12.23*	4.48	-5.78*	0.37
	(1.90)	(2.24)	(5.23)	(6.58)	(6.87)	(4.59)	(2.81)	(7.95)
Post Q-low	-0.89	8.72***	9.93*	3.32	7.57	-2.70	1.33	-1.69
	(1.57)	(2.59)	(5.69)	(7.50)	(6.51)	(2.86)	(4.22)	(8.07)
Post Q-medium	-1.18	2.72	-0.75	-5.66	-4.00	-3.45	-0.68	-4.80
	(1.64)	(2.01)	(5.40)	(6.91)	(6.59)	(2.92)	(3.68)	(6.88)
Post Q-high	-2.74	3.02	-5.36	-7.47	-5.24	0.27	0.56	-0.53
	(1.60)	(1.94)	(3.40)	(5.57)	(5.90)	(3.71)	(2.60)	(7.39)
Mother's age	-0.16*	0.20*	0.46**	0.21	0.35	0.21	0.01	0.01
	(0.09)	(0.10)	(0.19)	(0.18)	(0.21)	(0.14)	(0.08)	(0.15)
Mother's education	0.19	0.26***	0.85***	-0.21	0.41**	0.47*	0.29**	0.10
	(0.17)	(0.08)	(0.25)	(0.30)	(0.19)	(0.24)	(0.13)	(0.26)
Hhold size	0.21	0.16	0.30	0.57	-0.60	0.65	-0.10	-0.12
	(0.25)	(0.21)	(0.47)	(0.46)	(0.67)	(0.48)	(0.23)	(0.69)
N children $\leq$ 5 years	-0.25	-1.47**	-2.00*	-3.48***	-2.04	-3.19***	-0.74	0.48
	(0.42)	(0.56)	(1.13)	(0.97)	(1.33)	(0.62)	(0.72)	(1.08)
N births	-0.08	-0.37	-0.59	-0.83	0.10	-0.62	0.73**	-0.02
	(0.27)	(0.30)	(0.53)	(0.60)	(0.44)	(0.45)	(0.31)	(0.44)
Wealth Quintile	0.46	0.29	1.89**	0.90	1.86***	0.76	-0.04	0.77
	(0.30)	(0.34)	(0.75)	(0.91)	(0.58)	(0.52)	(0.24)	(0.68)
Married/cohabitating	3.72***	-0.44	0.61	-1.56	-0.97	-2.11	0.49	-0.19
	(1.28)	(0.88)	(2.42)	(2.14)	(2.17)	(1.73)	(1.10)	(1.89)
Insured	1.24*	0.39	1.61	2.80*	1.41	0.67	-0.65	3.38**
	(0.67)	(0.63)	(1.58)	(1.42)	(1.62)	(1.11)	(0.95)	(1.57)
Child's age								-0.42
								(1.42)
Male child								2.31**
								(1.08)
Birth year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
District FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N†	3,995	3,830	3,829	3,827	3,959	3,803	3,823	3,856
R2	0.02	0.02	0.04	0.21	0.02	0.04	0.01	0.01

* p<0.10, ** p<0.05, *** p<0.01 OLS models, standard errors clustered within districts. Scaled by 100 for exposition. † Details on varying reference population and covariate sets in Data section and Appendix. Models on child's vitamin A also control for child's age and gender.

Table B.5: Effect of P4P on outcomes (percentage point changes)

		Adverse events				
	(1)	(2)	(3)	(4)	(5)	(6)
	Breastfed at least 6 months	Any vision problems in last pregnancy	Night vision problems in last pregnancy	Maternal anemia in current pregnancy	Child anemia	Child infection in past 2wks
A. Average effect						
Treat*Post	1.04	1.04	1.13	8.11	3.34	4.47
	(1.49)	(3.71)	(1.53)	(16.16)	(7.49)	(6.19)
Post	1.72	-1.87	-1.95	-7.32	-28.55***	-17.77***
	(1.96)	(2.72)	(1.71)	(13.66)	(5.46)	(5.40)
Mother's age	0.06	0.13	0.07	1.47***	-0.45**	0.04
	(0.09)	(0.14)	(0.08)	(0.42)	(0.21)	(0.18)
Mother's education	0.02	0.11	-0.10	0.35	-0.05	-0.36
	(0.16)	(0.27)	(0.13)	(1.29)	(0.21)	(0.24)
Hhold size	-0.06	0.32	-0.10	-2.86	-0.49	0.35
	(0.56)	(0.37)	(0.23)	(2.27)	(0.54)	(0.61)
N children $\leq$ 5 years	8.60***	-1.11	0.05	6.53*	-0.19	-0.94
	(0.93)	(0.65)	(0.38)	(3.33)	(1.61)	(1.64)
N births	-0.58	0.15	0.20	-0.87	1.38*	-0.87
	(0.37)	(0.62)	(0.28)	(1.67)	(0.70)	(0.64)
Wealth Quintile	0.03	0.01	-0.09	0.03	-1.10**	-0.72
	(0.23)	(0.36)	(0.19)	(1.94)	(0.51)	(0.90)
Married/cohabitating	2.58**	0.54	0.60	6.58	-3.21	-2.41
	(1.04)	(1.35)	(0.78)	(10.36)	(2.16)	(2.20)
Insured	-1.52	0.17	0.84	0.59	-5.67***	1.00
	(1.02)	(1.54)	(0.72)	(4.50)	(1.96)	(1.63)
Birth year FE	Yes	Yes	Yes	Yes	Yes	Yes
District FE	Yes	Yes	Yes	Yes	Yes	Yes
N†	5,430	3,972	3,972	492	3,734	5,175
R2	0.06	0.01	0.01	0.02	0.08	0.03
Pre-period mean (%)	89.59	12.67	2.92	21.21	51.26	43.00
Incentive in USD						

* p<0.10, ** p<0.05, *** p<0.01 OLS models, standard errors clustered within districts. Scaled by 100 for exposition. † Details on varying reference population and covariate sets in Data section and Appendix. Models on child's anemia and infection also control for child's age and gender.

Table B.6: Effect of P4P on outcomes (percentage point changes)

		Adverse events				
	(1)	(2)	(3)	(4)	(5)	(6)
	Breastfed at least 6 months	Any vision problems in last pregnancy	Night vision problems in last pregnancy	Maternal anemia in current pregnancy	Child anemia	Child infection in past 2wks
B. Effect by baseline quality Q (results for covariates omitted)						
Treat*Post Q-low	-0.21	-1.71	1.26	25.75	8.30	10.56**
	(2.38)	(4.69)	(2.18)	(15.04)	(12.71)	(4.02)
Treat*Post Q-medium	4.95*	5.50	0.41	-18.78	3.23	17.01*
	(2.72)	(4.07)	(2.36)	(27.68)	(8.03)	(8.25)
Treat*Post Q-high	-2.22	1.00	2.97	-4.42	-1.87	-14.74
	(4.23)	(4.20)	(2.45)	(10.35)	(9.01)	(11.36)
Post Q-low	1.93	3.21	-0.73	-18.48	-39.02***	-21.72***
	(2.19)	(2.42)	(1.60)	(13.89)	(8.09)	(2.09)
Post Q-medium	0.37	-6.97**	-3.30	14.19	-24.14***	-25.30***
	(3.21)	(3.26)	(2.49)	(25.33)	(7.39)	(7.63)
Post Q-high	2.68	-3.23	-2.31	2.49	-23.07***	-7.48
	(3.06)	(3.21)	(2.26)	(5.57)	(6.61)	(9.10)
Mother's age	0.06	0.13	0.07	1.64***	-0.44**	0.04
	(0.09)	(0.14)	(0.08)	(0.54)	(0.21)	(0.18)
Mother's education	0.02	0.11	-0.10	0.23	-0.02	-0.36
	(0.16)	(0.26)	(0.12)	(1.27)	(0.21)	(0.24)
Hhold size	-0.06	0.31	-0.10	-2.94	-0.44	0.36
	(0.56)	(0.37)	(0.23)	(2.34)	(0.53)	(0.62)
N children $\leq$ 5 years	8.61***	-1.07	0.04	6.92*	-0.24	-0.99
	(0.93)	(0.65)	(0.38)	(3.36)	(1.61)	(1.62)
N births	-0.57	0.15	0.19	-1.30	1.33*	-0.86
	(0.37)	(0.62)	(0.28)	(1.89)	(0.69)	(0.64)
Wealth Quintile	0.05	-0.03	-0.13	-0.01	-0.90*	-0.60
	(0.22)	(0.36)	(0.18)	(2.13)	(0.45)	(0.91)
Married/cohabitating	2.61**	0.39	0.53	7.05	-2.96	-2.58
	(1.05)	(1.35)	(0.81)	(11.03)	(2.12)	(2.20)
Insured	-1.56	0.15	0.88	1.35	-5.64**	0.89
	(1.01)	(1.52)	(0.71)	(4.31)	(1.95)	(1.65)
Birth year FE	Yes	Yes	Yes	Yes	Yes	Yes
District FE	Yes	Yes	Yes	Yes	Yes	Yes
N†	5,430	3,972	3,972	492	3,734	5,175
R2	0.06	0.01	0.01	0.03	0.07	0.03

* p<0.10, ** p<0.05, *** p<0.01 OLS models, standard errors clustered within districts. Scaled by 100 for exposition. † Details on varying reference population and covariate sets in Data section and Appendix. Models on child's anemia and infection also control for child's age and gender.

The Consumption Smoothing Benefits of Health Insurance: Evidence from the Mexican Health Care Evaluation

Heidi Schramm

Draft Date: December 6, 2012

Abstract. I study the consumption smoothing benefit of health insurance in rural Mexico. Major illness is one of the largest and least predictable shocks to the economic wellbeing of families, inducing both high health care expenditures and loss of labor earnings. In rural areas, households smooth consumption via informal insurance mechanisms such as selling assets, drawing from savings, and borrowing from friends. However, these coping mechanisms may not provide full consumption insurance. In 2004, the Mexican government introduced one of the largest expansions of public health insurance in the world. The reform extended HI coverage to the families of all informal sector workers and significantly decreased out-of-pocket medical expenditures among beneficiaries. I use data from an experimental evaluation of the program to explore the differential effect of consumption smoothing over health shocks by HI status. I find that HI offsets illness-induced declines in consumption by over 50%. There are substantially larger effects for Oportunidades households, the particularly vulnerable segment of the population affiliated with the conditional cash transfer program, for whom HI offsets illness-induced declines in consumption by 75%.

1. Introduction

Families in developing countries are exposed to substantial financial risks due to major illness. Poor health induces high medical care costs and potential losses in labor income due to declines in labor supply and productivity. In rural areas of developing countries, credit markets are often incomplete and many individuals lack formal health and disability insurance. Households rely on informal insurance mechanisms, such as self-insurance through saving and inclusion in social risk networks, to smooth consumption over health shocks. However, these coping mechanisms may not provide full consumption insurance.

Many developing countries are considering social insurance to help insure vulnerable families against the economic costs of illness.[1] In 2004, the Mexican government introduced one of the largest expansions of public health insurance (HI) in the world. The reform extended HI coverage to the families of all informal sector workers through the introduction of Seguro Popular en Salud (SP). SP is a voluntary option available to the uninsured households who comprised about half of the Mexican population at the time of implementation. Previous research finds no significant short-term effects on the health care demand and health outcomes of SP beneficiaries.[2] However, SP has been found to significantly reduce out-of-pocket medical expenditures (OOP).

My research expands upon these findings by analyzing one of the non-health benefits associated with social insurance: consumption smoothing. Coupled with the fact that SP does not affect health care demand and health, the decline in OOP implies that SP affects household resources directly through the budget constraint in the short-term. If households are fully insured against health shocks through informal mechanisms, then social insurance may simply crowd-out informal insurance mechanisms. Alternatively, households may not be fully insured against illness and decrease consumption on non-medical goods. HI may moderate these non-medical consumption declines by decreasing OOP.

Few households pay the annual insurance premium associated with SP. Therefore, SP-affiliated households with low total medical expenditures may increase consumption: healthy households may decrease total income allocated toward precautionary savings in the face of future health uncertainty when they become insured. Thus, we may observe an increase in consumption even for families whose members' health does not change over the period of evaluation.

[1] Reforms are underway in several developing countries, including: China (Wagstaff et al., 2007), Colombia (Trujillo et al., 2005), Turkey (Agartan, 2005), Nicaragua (Thornton et. al., 2010) and Vietnam (Wagstaff, 2007).
[2] I review the previous findings on SP in Section 4.2.

Identification of the effects of health insurance on health and related outcomes is empirically difficult because individuals with and without HI differ in ways that are correlated with outcomes of interest. For example, an individual's unobserved propensity to purchase HI may be related to his underlying health profile, risk preferences, or inclusion in social risk-sharing networks. To circumvent endogeneity concerns, I use data from the Mexican Health Care Evaluation, a large-scale randomized field experiment that took place in 2005 in seven Mexican states. The experiment was designed to mimic the national roll-out of SP and consisted of an encouragement campaign, randomly assigned at the community level, which increased awareness of the program and encouraged affiliation with SP. While any uninsured individual was eligible to sign up for SP, including those in control areas, the program was not well-known when the experiment was implemented and affiliation was low. The encouragement campaign had a large and significant effect on individual HI status of 38.6 percentage points. I use exposure to the campaign, determined by community of residence, as a factor that exogenously affects binary HI status.

Consistent with previous research, I find that households are not fully insured against the economic costs of illness. Using detailed individual health data, I find that a full health deterioration of an adult household member lowers annualized per-capita-expenditure (PCE) by 14%. I implement an instrumental variables strategy to explore the differential effect of consumption smoothing over health shocks by HI status. My results indicate that HI offsets illness-induced declines in consumption by over 50%. I find substantially larger effects for Oportunidades households, the particularly vulnerable segment of the population affiliated with the conditional cash transfer program, for which HI offsets illness-induced declines in consumption by 75%.

There is little existing research that assesses the extent to which HI enables households to smooth consumption over changes in health, in the context of both developed and developing countries. The results of this paper have important policy implications. In the short term, the social insurance program did not affect several outcomes that were important to policymakers, such as health care utilization or health. However, the program did significantly reduce the incidence of catastrophic medical expenditure and provide a consumption-smoothing benefit to families who faced serious health shocks.[3] This is an important finding because previous evidence on public income transfers in Mexico suggests that private transfers could neutralize changes in public programs (Juarez, 2009). My empirical work indicates that households are not fully insured through

[3] A household has catastrophic medical expenditures when more than 30% of household spending is allocated toward health. In the data, 9% of households face catastrophic health expenditures at baseline.

informal mechanisms and HI provides a consumption-smoothing benefit, as opposed to simply crowding out informal insurance. However, HI does not fully offset the economic costs associated with illness and my estimates suggest that public disability insurance may provide further benefits by insuring households against the labor earnings losses associated with illness.

2. Related Literature

I contribute to two rich areas of literature: consumption smoothing and health insurance (HI). There is little empirical work that assesses the extent to which HI helps households smooth consumption over health shocks, in the context of either developed or developing countries.

2.1 Consumption Smoothing

Research on consumption smoothing is comprised of two strands of literature: that which investigates the strategies that households use to smooth consumption, and that which evaluates the effectiveness of these strategies. This paper is most closely aligned with the latter branch of literature. Drawing from the seminal work of Townsend (1994), much of this second strand of literature evaluates the efficacy of consumption smoothing strategies by analyzing the extent of consumption smoothing in the context of the theory of full insurance. The theory of full insurance posits that households within the same community fully share the risk of idiosyncratic shocks: growth in household consumption does not depend on changes in household resources that are uncorrelated with shifts in preferences after controlling for changes in aggregate community resources.

While much of this literature is devoted to analyzing consumption smoothing over income shocks, there are several studies that specifically examine consumption smoothing over health shocks. Townsend (1995) finds no effect of illness, as measured as the percentage of the year that an adult male is sick, on consumption in the context of India. Kochar (1995) finds that illness to male household members, as measured by an indicator of loss of work due to illness, lowers wage income and increases informal borrowing during peak periods of the agricultural cycle but has no effect during off-peak periods or for females. These early studies suggest that households are able to insure health shocks relatively well. However, as noted by Gertler and Gruber (2002), the health measures used in these studies may represent small anticipated changes in health and not the kind of large unexpected health shocks that are difficult to insure. It is precisely this type of health shock for which we would expect to see large benefits of health insurance.

The research most closely related to my work is that by Gertler and Gruber (2002), which analyzes health shocks in Indonesia. They find imperfect insurance of consumption over illness: a complete deterioration of the household head's health causes a 20% decrease in annualized per-capita consumption. The health measure they use is an index comprised of an individuals' ability to perform activities of daily living (ADLs). I use a similar index in my empirical work. In the context of both developed and developing countries, ADLs have proven to be a valid measure of health and may distinguish serious health problems (Stewart et al., 1990; Strauss et al., 1993).[4] Gertler, Levine, and Moretti (2009) test whether access to microfinance programs, as measured by geographical proximity to a microfinance institution, helps Indonesian families smooth consumption over illness. Their findings indicate that access to financial institutions assuages the effects of health shocks: for households located within 1 kilometer of a financial institution, negative health shocks have substantially smaller effects on consumption than for households living 10 kilometers or more from a financial institution. Implementing the empirical strategy of Gertler and Gruber (2002), I find results of a comparable magnitude in the context of Mexico. I extend the analysis to assess the extent to which HI moderates the decrease in consumption due to health deteriorations.

Several researchers modify the original full insurance model to study risk sharing over health shocks *within* the household and risk sharing with overlapping networks. Dercon and Krishnan (2000) investigate risk sharing within households in the context of Ethiopia. Their findings indicate that in the majority of households, full risk sharing occurs. However, in poor southern Ethiopian households, risk sharing does not occur when the illness shocks are to women. Using data on insurance networks in Tanzania, De Weerdt and Dercon (2006) test whether full risk sharing occurs within overlapping networks. They find full insurance of food consumption at the village level and full insurance of non-food consumption at the level of networks.

Research specific to consumption smoothing in the context of Mexico does not explicitly study health shocks. However, it does inform the strategies that Mexican households use to cope with adverse events. Ruiz (2011) studies the effect of expanded credit access on savings and consumption smoothing across income shocks. She finds that households use savings as a buffer on income fluctuations and that savings decline once formal credit is available. Hamoudi (2010)

[4] The only paper to account for the endogeneity of health shocks is Mohanan (forthcoming Review of Economics and Statistics). Mohanan treats bus accident injuries in India as exogenous health shocks and, using propensity score matching, finds full consumption insurance over small health changes. These results are consistent with the previous literature that does not account for the endogeneity of health shocks.

analyzes the role that risk aversion plays in household partition and within-family consumption smoothing in rural Mexico. He finds that household partition takes the form of diversification in production - split-off households are differentially selected into formal market employment and origin households in agricultural or non-agricultural self-employment- and families continue to share resources after partition. Juarez (2009) estimates the effect of an exogenous increase in the income of elderly individuals, due to a 2001 demogrant in Mexico City, and finds almost complete crowding out of private transfers. Her results highlight the importance of my work because they suggest that private transfers could neutralize changes in public transfers. Unfortunately, the data used here do not allow me to explicitly examine whether savings and inter-household transfers decline upon receipt of HI, but it is contextually critical to theoretically allow for such a possibility. My empirical work indicates that households are not fully insured through informal mechanisms and that HI does not completely crowd out private risk sharing arrangements. Furthermore, the use of savings and assets to finance medical expenditures is lower among the insured compared to the uninsured.

Finally, of particular relevance to my research on the heterogeneous effect of HI on Oportunidades households, Skoufias (2002) examines how PROGRESA, the rural conditional cash transfer program later extended to urban areas as Oportunidades, affects the consumption smoothing of households. His findings are of note for two reasons. First, he finds that informal insurance is incomplete with regard to income fluctuations, which my empirical work corroborates with respect to health shocks. Second, he finds that the cash transfer associated with PROGRESA allows households to better insulate their consumption from income shocks. My empirical findings on the heterogeneous effects of HI for Oportunidades households are consistent with his research: it is likely that HI assuages the effect of increased OOP due to negative health shocks, while the cash transfer protects the household against income loss due to illness.

2.2 Health Insurance

Recent interest has grown in providing social insurance to vulnerable populations as a means of increasing health care utilization and providing insurance against catastrophic medical expenditures. While there are many studies that compare the health and medical care utilization of the insured and uninsured, identifying the causal effect of HI on health related outcomes is empirically difficult because of the selection effects associated with health insurance. The insured and uninsured differ in ways that are potentially correlated with outcomes of interest, for example by risk preferences or inclusion in social risk sharing networks. There is limited rigorous empirical evidence on the effects

of HI on utilization, health expenditures, and health outcomes. Research on the non-health benefits associated with social insurance, such as consumption smoothing and reduction in out of pocket medical expenditure risk, is even thinner.

Evidence on the effect of HI on medical care utilization and medical expenditures in developing countries is mixed. Using data from an experimental evaluation of a voluntary HI program for informal sector workers in Nicaragua, Thornton et al. (2010) find that HI decreases OOP and does not increase health-care utilization. King et al. (2007) use data from their experimental evaluation of the Mexican program I consider here and find similar results to the Nicaragua study.[5] Trujillo et al. (2005) use propensity score matching to measure the effect of a Colombian health insurance subsidy on medical care utilization of low-income families and find that Colombia's subsidized insurance program increased medical care utilization among the poor and uninsured. Finally, Brugiavini and Pace (2010) find that the introduction of the Ghanaian National Health Insurance Scheme increases medical care utilization but has little effect on OOP.

There is a substantial body of work that investigates the effects of the social insurance programs Medicare and Medicaid in the U.S. Using a randomized lottery research design, Finkelstein et al. (2012) find that individuals in their treatment group, those who had the opportunity to apply for Medicaid, are more likely to have health insurance, have higher medical care utilization, and lower OOP and medical debt. Finkelstein and McKnight (2008) study the impact of the introduction of Medicare on elderly mortality and out-of-pocket medical expenditure risk. While they find there is no effect on mortality over the first 10 years of the program, they estimate that the introduction of Medicare was associated with a 40% decrease in OOP for the top quartile of the medical expenditure distribution.

This paper contributes to that small body of work that examines the non-health benefits of social insurance. An entirely unexplored area of research is that which assesses the extent to which HI affects the different mechanisms through which health may affect consumption. Health insurance may mediate the effect of health on consumption by modifying the effect of health on medical expenditures, labor earnings, savings, and transfers from risk-sharing networks. Unfortunately, the data used in this paper do not allow me to explore these mechanisms in detail.

[5] I discuss research directly related to the social insurance program Seguro Popular in Section 4.2.

3. Model

HI may affect consumption through three channels: medical expenditures, informal insurance and labor earnings. HI may operate directly through the budget constraint (BC) by decreasing OOP during illness. HI may also affect the BC via informal insurance mechanisms, such as precautionary savings and transfers from others. For example, the uninsured may save more due to uncertainty over future health or receive more transfers through social networks when sick. HI may increase health care utilization, and thus health, and result in higher income through health-induced increases in productivity or work hours. This is unlikely given that health care utilization and health are *not* affected by HI over the evaluation period. Thus, I assume HI operates exclusively through the BC.[6]

I develop a 3-period model to describe the relationship between health-induced changes in consumption and HI. To study this relationship, I focus on how the change in consumption due to health changes, between periods 1 and 2, varies by HI. I include a third period to capture forward-looking behavior in period 2. I allow the effect of HI to vary depending on the direction of health changes; that is, whether health deteriorates or improves between periods. This is because the optimal consumption and savings demand functions are slightly different depending on initial resources, which depends on initial health.

3.1 Health Deteriorations

There are 2 states: healthy and sick. In period 1, the individual is healthy. He has earned income w to finance consumption and faces no OOP. He can set aside savings s_1 to be used in period 2. He faces risk $(1-p)$ of becoming sick between periods. I assume the illness probability is exogenous.[7]

Sickness causes net income to fall to kw, where $0 \leq k \leq 1$, and OOP $(1-b)m$ that are proportional to the cost of treatment, where m is the total cost of treatment and b is HI coverage.[8] There are two additional sources of income to finance consumption: savings and transfers t from others. I assume transfers are gifts and, following Gruber (2000), model them as a function of net income in the sick state: $t = f(kw-(1-b)m)$. The individual does not receive transfers when healthy. In period 2, he allocates net income across consumption and savings s_2 to be used in period 3. In period 3, he consumes income and savings. He maximizes the 3-period utility function:

[6] Because the earnings channel is likely to operate in the long-term, my estimates are limited to the short-run.

[7] I assume the probability of illness is not a function of HI. This is a reasonable assumption for the short evaluation period and is supported in my data. In addition, I assume the illness probability is independent of preferences that potentially affect consumption-smoothing ability.

[8] To facilitate analysis, I do not allow medical expenditures to be a choice. I allow HI to be continuous between 0 and 1, where b=1 if the individual has 100% coverage. Empirically I do not observe this and treat HI coverage as dichotomous.

$$(1)\quad U = U(w - s_1) + \frac{1}{1+\delta}\Big(pU\big(w + (1+r)s_1 - s_2(h_2 = 1)\big) + (1-p)U(kw + (1+r)s_1 - s_2(h_2 = 0) - (1-b)m + t)\Big)$$
$$+ \left(\frac{1}{1+\delta}\right)^2 \Big(pU\big(w + (1+r)s_2(h_2 = 1)\big) + pU\big(w + (1+r)s_2(h_2 = 0)\big)$$
$$+ (1-p)U(kw + (1+r)s_2(h_2 = 1) - (1-b)m + t) + (1-p)U(kw + (1+r)s_2(h_2 = 0) - (1-b)m + t)\Big)$$

where δ is the discount rate and r is the rate of return on savings.[9] Period 2 savings s_2 depend on health in period 2, where $h_2 = 1$ means the individual is healthy.

Maximization of expected utility yields optimal savings $s_1{}^*(w, k, p, b, m, t)$ and $s_2{}^*(w, k, p, b, m, t)$ and resulting per-period consumption $c_i^*(w, k, p, b, m, t)$ for $i = 1,2,$ and 3, conditional on realized health in periods 2 and 3. I derive the consumption-smoothing effect of HI over illness by focusing on the change in consumption between periods 1 and 2, where the individual becomes sick in period 2. I assume no state dependence: the marginal utility of consumption *evaluated at the same level of consumption* is the same regardless of health state.[10] This allows me to write the change in consumption upon illness in terms of the BC:

$$(2)\quad \Delta c = c_2^* - c_1^* = (2+r)s_1{}^*(w, k, p, b, m, t) + t - (1-k)w - (1-b)m - s_2{}^*(w, k, p, b, m, t | h_2 = 0)$$

The first term is the increase in consumption associated with drawing from period 1 savings. The second term is the increase in consumption associated with state-contingent transfers from others. The third term is the loss in consumption due to the labor earnings cost of illness. The fourth term is the loss in consumption due to OOP. The last term is the loss in consumption due to savings for the next period. If $\Delta c < 0$, households experience a decrease in consumption due to health declines.

To study the extent to which HI b offsets illness-induced changes in consumption Δc, the comparative static of interest is $\frac{d\Delta c}{db}$. While medical expenditures may be fully insured ($b = 1$), individuals do not have insurance against the income loss associated with illness. Unless individuals are fully insured through informal mechanisms, I expect HI to moderate, but not fully offset, declines in consumption. HI has two effects on consumption changes: the direct effect on OOP and indirect effects on the informal mechanisms. The comparative static can be written as:

$$(3)\quad \frac{d\Delta c}{db} = (2+r)\underset{(?)}{\frac{ds_1{}^*}{db}} + \underset{(-)}{\frac{dt}{db}} + \underset{(+)}{m} - \underset{(?)}{\frac{ds_2{}^*}{db}}$$

[9] I assume the realized health state in period 2 does not affect the probability of illness in period 3. Relaxing this assumption does not change my comparative statics because I look at the differential effect of changes in health on changes in consumption across HI status, and therefore compare individuals who have similar health changes and face a similar illness probability.

[10] Allowing for state dependence adds a term to equation (2) that represents the change in consumption purely due to state dependence. I discuss the implications of state dependence in Section 8.

The first term is ambiguous due to 2 competing effects. As insurance increases there is less need for precautionary savings. However, there is a positive indirect effect of b on savings through the transfer function: transfers decrease as b increases, increasing savings. The second term is negative: transfer income decreases as HI increases. The third term is the direct effect of HI on OOP, which can instead be allocated to non-medical consumption. Finally, optimal period 2 savings respond ambiguously to HI. There is a negative effect through period 1 precautionary savings: an increase in coverage causes a decrease in s_1 and therefore there are fewer resources available (for saving and consuming) in period 2. There is also a negative effect through period 2 precautionary savings: an increase in coverage causes a decrease in the need for precautionary savings for period 3. However, the indirect effect of b on savings through the transfer function is positive. The aggregate effect of HI on illness-induced consumption changes is ambiguous and depends on the relative magnitudes of total medical costs and the indirect effects of HI on informal mechanisms.

3.2 Health Improvements

To examine the effect of health improvements on consumption by HI, I use the same model with one adjustment. I assume the individual is sick in period 1 and consider consumption changes across periods 1 and 2, where the individual becomes healthy. Similar to equation (2), the change in consumption upon a health improvement is:[11]

$$\Delta\hat{c} = \hat{c}_2^* - \hat{c}_1^* = (2+r)\hat{s}_1{}^*(w,k,p,b,m,t) - t + (1-k)w + (1-b)m - \hat{s}_2{}^*(w,k,p,b,m,t) \quad (4)$$

If $\Delta\hat{c} > 0$, health improvements increase consumption. The differential effect of health improvements on consumption by HI status is not obvious. Health improvements raise labor earnings and reduce medical expenditures. If the individual has HI, consumption will change little due to decreased medical expenditures. Alternatively, the uninsured may exhibit a large increase in consumption, especially if their first period OOP were not substantially covered by informal mechanisms. Labor income will increase regardless of HI. Whether this increase in labor income is allocated toward consumption or savings for the following period *does* depend on HI. Individuals with HI may allocate a substantial portion of their increased labor earnings toward current period consumption, while individuals without HI may replenish their savings and exhibit little change in consumption. The comparative static $\frac{d\Delta\hat{c}}{db}$ represents how health improvement-induced changes in consumption respond to HI and can be written as:

[11] I denote the Marshallian demand functions from the health improvement problem different from those in the health deterioration problem because these functions are not identical ($\hat{c}_1^*$ instead of c_2^*).

$$(5) \qquad \frac{d\Delta\hat{c}}{db} = (2+r)\frac{d\hat{s}_1^{\,*}}{db} - \frac{dt}{db} - m - \frac{d\hat{s}_2^{\,*}}{db}$$

$$(?) \quad (-) \quad (+) \quad (?)$$

The sign of the first term is ambiguous. As insurance increases, more period 1 income is available to allocate to $\hat{s}_1$. However, as b increases the need for precautionary savings in period 2 declines. The indirect effect of b on savings through the transfer function is positive. The sign of the last term is also ambiguous. First, the direct effect of b on period 1 savings is positive: HI increases period 1 resources and thus s_1 resulting in more period 2 resources to be allocated toward s_2. Second, the precautionary motive on period 1 savings is negative: HI decreases the need for s_1 and therefore the resources available for s_2 decrease. Third, the precautionary savings motive on period 2 savings is negative: HI causes a decrease in the need for savings in period 3. And finally, the effect of b on s_2 via transfers is positive. The aggregate effect of HI on health improvement-induced changes in consumption is ambiguous, depending on the magnitude of m and the effect of HI on savings.

4. Background and Program Description

The Mexican health care system is comprised of several vertically integrated social security institutions.[12] The Mexican Social Security Institute (IMSS) is the largest institution and provides social security benefits to formally employed, salaried private-sector workers and their families. The second largest institution, Institute of Security and Social Services for State Workers (ISSSTE), offers benefits to government workers. Several smaller institutions serve employees of the navy (SEMAR), the national defense ministry (SEDENA), and the state-owned oil monopoly company (PEMEX).

The Mexican social security system (SS) is mandatory for workers in the formal sector and provides health care benefits for beneficiaries and their families. The various institutions employ salaried doctors and operate their own hospital facilities and clinics. The institutions are financed by a combination of employer, employee, and federal government contributions. Health-care benefits include all services provided at each SS facility: acute treatment and outpatient care in hospitals and ambulatory clinics and prescription drug coverage.[13] In 2005, around half of the population had access to health insurance through SS.

Three percent of the Mexican population purchases private health insurance. The private insurance market is highly concentrated; the two largest companies serve 50% of the market. There

[12] This section draws from the OECD Review of Health Systems: Mexico (2005).

[13] There is not a comprehensive list of services aside from the generic entitlement to health-care coverage indicated in the 1984 General Health Law. For more detail see OECD (2005) and Homedes and Ugalde (2009).

are two policy types: catastrophic medical expenditure and HMO-type policies. The high premiums associated with private insurance are a financial barrier for the majority of Mexican population (OECD 2005).

The uninsured can purchase medical care in public clinics and hospitals or in the large, unregulated private market.[14] Health care in public clinics, operated by the Ministry of Health and State Health Services, is subject to payment of income-related user fees. While care in public clinics is often below full cost, private out-of-pocket medical expenditures accounted for 55% of all health-related spending in Mexico in 2005 (OECD 2005).

4.1 Seguro Popular

In January 2004, the federal government passed legislation that introduced the System for Social Protection in Health (Sistema de Proteccion Social en Salud, SPSS). As detailed by Julio Frenk, the Minister of Health from 2000-2006, in addition to increasing health care utilization and improving health outcomes, one of the main goals of the program was to eliminate catastrophic health expenditures for informal sector workers, which placed uninsured families at a high risk of impoverishment (Frenk 2006, Frenk et al. 2006).[15] The SPSS reforms were legally mandated to be phased in over seven years and contain both supply- and demand-side initiatives.

At the center of the reforms was Seguro Popular en Salud (SP). SP was intended to be a wealth transfer program, rather than an insurance scheme, because premiums are not risk adjusted and not set to break even. However, economically SP is a voluntary health insurance option available to all citizens not covered by SS. Individuals must formally affiliate with SP and pay an annual income-based fee.[16] Households that participate in Oportunidades, Mexico's conditional cash transfer program, are automatically enrolled. SP provides coverage for over 250 health interventions and more than 300 pharmaceuticals: the services covered by SP treat illnesses accountable for 95% of the disease burden in Mexico (King et al. 2007).

Due to financial constraints, SP was designed to be gradually rolled out and only to communities whose health facilities met a set of minimum infrastructure and staffing requirements. While individuals living in non-SP communities could travel to the nearest SP community to affiliate

[14] Affiliates of the Oportunidades conditional cash transfer poverty alleviation program receive free basic preventative health care at State Health Services or IMSS-Oportunidades facilities.(OECD 2005)
[15] Homedes and Ugalde (2009) detail the political history of Mexican health reforms.
[16] The annual fee is not enforced; in the data only 6.4% of affiliates report paying a premium. As discussed by Lakin (2010), the public health system is severely underfinanced at present.

and receive care, there is little evidence that this occurs during the time under consideration, when awareness of the program was low (King et al., 2007).

4.2 Previous Research on Program Effects

The first non-experimental studies on program effects were published in *The Lancet* in 2006, in a series of articles on the reform. Gakido et al. (2006) find that SP affiliation successfully reaches the poor, marginalized communities, federal non-social security expenditure per person increased by 38% between 2000 and 2005, equity of public health expenditure across states improved, SP affiliates used more inpatient and outpatient services compared to uninsured people, and catastrophic expenditures for SP affiliates were lower than for the uninsured. Knaul et al. (2006) present 15-year trends on the incidence of catastrophic health expenditures and suggest that the recent downward trend is attributable to SP.

There was substantial variation in the introduction and program penetration of SP across regions. As noted by King et al. (2007), political considerations played a large role in determining regional variation and raised concerns about identification strategies based on non-experimental variation. Because of these concerns, the Mexican Ministry of Health commissioned the Mexican Health Care Evaluation in 2005. King et al. (2007) details the political context and experimental evaluation. The experimental design, randomization within matched health cluster pairs, ensured that the benefits of the randomization were not lost if a health cluster dropped out of the experiment before the evaluation. This was a serious concern because a federal election was held in 2006, the year of the evaluation.[17]

King et al. (2009) is the first paper to examine program effects using the experimental data. They estimate intent-to-treat and complier average causal effects (CACE) non-parametrically. SP reduced by 23% (55% CACE) the proportion of households experiencing catastrophic expenditures compared to baseline. They find no significant effects on the use of health care services and no effect of SP on various health measures, both self-reported and objective.

Spenkuch (2011) and Barofsky (2011) use the experimental data to study moral hazard and provide a cost-benefit analysis of the program, respectively. Spenkuch (2011) measures determinants of insurance enrollment by regressing HI status in the follow-up survey on the randomly assigned treatment, baseline controls, variables that indicate riskiness (self-reported health and objective

[17] I discuss the experimental evaluation and data in the following section.

health measures at baseline) and allows treatment effects to vary based on the vector of riskiness variables. He finds that individuals who assess their own health as "good" or "very good" at baseline are 5.5 and 6.9 percentage points less likely to take up HI than individuals who assess their health as "fair", "bad" or "very bad". Surprisingly, there is no indication of selection on objective measures of health, such as BMI, cholesterol or blood pressure. To test for ex ante moral hazard, Spenkuch uses the data on utilization of preventative medical care, such as whether an individual had a flu shot or a mammogram. Both ITT and LATE estimates indicate that HI coverage induced individuals to engage in less self-protection. However, these findings have limited implications for my empirical work because there is no effect on actual health outcomes.

Barofsky (2011) provides a cost-benefit analysis of SP and is the first research on SP to estimate one of the non-health benefits of insurance: reduced risk exposure. He calculates the social welfare gain associated with a decrease in risk premiums due to SP by comparing the out-of-pocket health expenditure distributions between treatment compliers in treatment and control groups during the follow-up survey, much in the spirit of Finkelstein and McKnight (2008). He concludes that the benefits of the program only represent 62% of the cost.

There are several papers that use non-experimental data to study the effect of the program on utilization, health care spending, and health outcomes. All of these papers assume that the geographical variation in SP roll-out is exogenous. Knox (2008) exploits municipality-level variation in SP enrollments to analyze the effect of the program on health spending, health outcomes and labor supply. Her difference-in-difference estimates indicate that SP increases health care utilization, does not effect health spending or health outcomes, decreases labor supply for secondary workers, and increases hours worked for older adults.

Barros (2011) implements a triple-difference estimation strategy using repeated cross-sectional data to examine effects on the use of public and private providers, OOP, and labor force participation. He uses variation in state-level program intensity targets over time and across states between 2000 and 2006 to examine the effects on health and labor market outcomes. His estimates indicate that SP: induced beneficiaries to shift from private to public providers, decreased out-of-pocket health expenditures, did not have an effect on health outcomes, and did not affect labor force participation. Barros hypothesizes that the care provided under SP is low enough that it did

not encourage movement to the informal sector. However, this is not supported in the experimental data I use here (King et al. 2009).

In addition to the findings mentioned above in Knox (2008) and Barros (2011), there are two papers that study the labor market effects of the program. Arias et al. (2010) is primarily concerned with assessing the effect of monopoly and regulation on the size of the informal labor market. They conclude that the introduction of SP did not promote growth of the informal sector using graphical analysis of national trends in illegal salaried employment. Effects of the program on informal sector employment are also analyzed in Azuara and Marinescu (2011). Their findings indicate that SP has a negative and insignificant effect on informality. When they limit their sample to less educated workers (9 years of schooling and less), they find that SP significantly increases informality: less educated workers are .8 percentage points more likely to work informally after the introduction of SP. This is a very small effect (less than 2%) given that half the labor force in this group is informal. Their results suggest that very few workers choose labor sector on the basis of the availability of health benefits.

The above findings inform my theoretical and empirical work. First, because there is no effect of SP on health care utilization or health outcomes, I assume that HI affects consumption growth purely through its effect on OOP and informal insurance mechanisms. If HI affected health transitions, both my simplified model and empirical strategy would not be appropriate. Theoretically, HI would modify the effect of health on consumption both through the budget constraint and through its effect on labor earnings via health outcomes. I assume HI operates exclusively through the budget constraint in my model. If HI affected health transitions, this would be empirically problematic because HI would affect the independent variable across which I examine differential effects. Second, I do not model individual labor market decisions. If workers made employment decisions based on insurance benefits we may observe a correlation between health insurance status and income, as dictated by the theory of compensating differentials.

5. Mexican Health Care Evaluation & Data

I use data from the experimental evaluation of Seguro Popular en Salud (SP). The treatment was a community-level encouragement design that mimicked the national roll-out of SP. The treatment occurred in August 2005 before many households had affiliated with SP. In treated communities, affiliation with SP was encouraged via local media campaigns. Information stands, where individuals

could sign up for the program, were set up throughout treatment communities to decrease the travel costs of affiliation. In control areas, households affiliated at a local health clinic.

The evaluation team used a randomized cluster matched pair design: they partitioned Mexico into 12,284 health clusters, defined as the catchment area of an existing health clinic, and randomly assigned treatment within 74 matched cluster pairs in 7 Mexican states. Figure 1 shows a map of participating states. 90% of the pairs are in rural areas and they were matched on characteristics from nationally representative data.[18] While any uninsured individual was eligible to sign up for SP, including those in control areas, the program was not well-known when the experiment was implemented. Figure 2 shows the proportion of individuals with any type of HI before and after treatment by treatment and control groups. The encouragement campaign had a large and significant effect on individual HI status of 38.6 percentage points. I use exposure to the campaign, determined by community of residence, as a factor that exogenously affects HI status.

The data consists of baseline and 10-month follow-up surveys for over 30,000 households across 50 matched clusters.[19] The baseline survey was conducted at the time of random assignment in August 2005. The surveys collect household demographics and spending data from the household head, and detailed individual health data from a randomly selected adult. The spending data is collected across a variety of medical and non-medical goods, such as: hospitalizations, outpatient visits, food, and education. I observe individual HI status, HI type, and residence by treatment group. The data does not include income, earnings, savings, or transfers. This is a data limitation that hinders evaluation of the various mechanisms through which HI moderates health-induced changes in consumption. Descriptive statistics are provided in Table 1.[20] The sample is comprised of primarily rural households and nearly half are Oportunidades recipients.[21] Of individuals for whom I have detailed health data, 64% are women, and individuals exhibit low levels of education and employment. I consider growth in per capita expenditure (PCE) for monthly food goods and growth

[18] The experiment intentionally over-sampled rural areas with a large proportion of uninsured individuals. The sample is otherwise nationally representative.

[19] The household surveys were limited to 50 matched clusters, out of 74, due to budget constraints.

[20] After removing observations with missing variables and trimming the consumption growth variables (the lower and upper 1 percent of the distribution), I have 24,956 observations.

[21] Oportunidades is a conditional cash transfer program for families with school-age children. The effect of HI on Oportunidades families is of particular interest because these households are in the lowest two income deciles, automatically enrolled in SP, and receive basic preventative health care and cash transfers through Oportunidades. I explore the effect of HI on Oportunidades households in section 7.1.

in annualized household PCE as dependent variables. Distributions of inflation-adjusted consumption growth are provided in Figures 3 and 4.

5.1 Health Variables

My independent variable of interest is health changes. For the randomly selected adult in the household for whom I have detailed health data, there are several potential variables to measure changes in health: objective measures (weight, height, blood pressure, and blood tests regarding blood sugar and diabetes), self-reported health status across 15 different categories, and indicators for previously diagnosed diseases (heart disease, hypertension, diabetes). Descriptive statistics of various health variables at baseline are in Table 2. As is evident in the data, Mexico's disease profile is increasingly moving toward that of a developed country: more than half of the sample is overweight and the average blood pressure is in the pre-hypertension or hypertension stage. Health measures for treatment and control groups are similar at baseline.

Following Gertler and Gruber (2002), I measure changes in health as the difference between a baseline and follow-up health index. For the randomly selected adult, I create a health index from answers to questions regarding level of ability in completing a variety of activities, where individuals rate their ability on a scale of 1 to 5.[22] Summary statistics of self-reported health variables at baseline are provided in Table 3. I sum the scores across the activities and compute the following:

$$Index_i = \frac{Score_i - Minimum\ Score}{Maximum\ Score - Minimum\ Score}$$

The index takes on a value of 1 if the individual reports no problems with the any activities and 0 if the individual cannot do any activities. Table 4 reports summary statistics for both baseline and changes in the index. The index changes for many individuals between surveys. 68.5% of adults report some limitation at baseline. Between surveys, 22% report deterioration and 53% report an improvement. On average, the size of the deteriorations and improvements are comparable. The smallest change is -.042, equivalent to moving from a difficulty of "none" to "mild" for one activity. The largest deterioration is -1, as some report no problems at baseline but "can't do" any activities later. The distribution of health index changes is provided in Figure 5.

5.2 Health Insurance

As discussed in Section 4, the majority of insured individuals receive HI through the labor market. There is no evidence that individuals choose labor sector based on the availability of health benefits

[22] Activities include: moving around, vigorous activity, self-care, grooming, cooking, and performing work.

(Azuara and Marinescu 2011, Arias et al., 2010). Table 5 shows that 21.5% of individuals are insured at baseline and of those insured, 64% have obligatory health insurance through the formal labor market. After the experiment, 42.8% of individuals have HI and 61.4% of individuals have insurance through SP. Table 6 verifies that the large increase in the insured population occurs primarily in treated areas: 40% of individuals residing in treatment areas have no HI at baseline but are covered at the time of the follow-up, compared to 10% in the control areas. There are more households with HI at baseline in the treatment group. As discussed by King et al. (2007), this is because the baseline survey was conducted as the encouragement campaign was starting. Because insurance coverage occurs 30 days after sign-up, this is unlikely to affect actual outcomes, but households in treatment areas are more likely to report that they have HI after signing up but before coverage begins.

One limitation of the data is that I do not observe total medical expenditures before insurance coverage, and thus cannot treat HI as a continuous variable. This is not a prohibitive concern because benefits under SP were designed to be similar to those provided to formal sector individuals, who make up over half of the insured population at baseline. I use a binary HI variable and do not differentiate between the various types of insurance in my empirical work.

6. Empirical Strategy

Identification of the effects of health insurance (HI) on health and related outcomes is empirically difficult because individuals with and without HI differ in ways that are correlated with outcomes of interest. For example, an individual's unobserved propensity to purchase HI may be related to his underlying health profile, risk preferences, or inclusion in social risk-sharing networks. As previously discussed, the majority of insured individuals have obligatory HI through the labor market and there is no evidence that individuals make labor market decisions based on insurance coverage. However, using experimental data further alleviates selection concerns: the data reduces potential omitted variable bias by ensuring that control and treatment groups are comprised of similar types. The previous section verifies that the treatment and comparison groups are similar on observable baseline characteristics.

The encouragement campaign exogenously affects the probability that an individual has HI. I use the campaign as an instrument for binary HI status and discuss the effect of the campaign in the following section. I then present results on consumption smoothing over illness, followed by my main specification: the differential effect of health changes on consumption growth by HI status. I also consider heterogeneous effects for Oportunidades households, households where the

household head is the randomly selected individual for whom I have detailed health data, and households where the individual is employed at baseline.

6.1 Effect of the Encouragement Campaign

The encouragement campaign was designed to increase awareness of the SP option and to decrease travel costs associated with program affiliation. I estimate the effect of being exposed to the encouragement campaign on the individual in household h and village v with the following regression equation:

$$(6) \qquad HI_{hv} = \alpha_v + \alpha_1 Campaign_v + X'_{hv}\alpha_2 + \varepsilon_{hv}$$

where HI is insurance status in the follow-up survey. The vector of household controls X_h includes: household size, urban/rural location, whether the household is indigenous, and Oportunidades enrollment. Because I use cluster match fixed effects to control for changes in potential determinants of both income and health in my main specification, I include them here. $Campaign$ is an indicator for whether the individual resides in a village with the randomly assigned encouragement campaign. The coefficient α_1 on $Campaign$ gives the average difference in means between the treatment and control group. As shown in Table 7, the encouragement campaign has a large and significant effect on individual HI status of 38.6 percentage points. The estimate is significant at the 1% level.[23]

In my main specification, I instrument for HI using the campaign as an instrumental variable. The exclusion restriction requires that the treatment, living in a cluster with the encouragement campaign, does not have a direct effect on the outcome under consideration (changes in consumption) and only affects consumption indirectly via HI. This restriction will not be met if the campaign directly affects an unobserved variable that is correlated with the outcome, such as prices or interest rates. Given the short period under consideration, it is unlikely that this occurred.

6.2 Effect of Illness on Consumption

I test whether household consumption is fully insured against the costs of illness using Gertler and Gruber's (2002) estimation strategy. This captures the reduced-form effect of health changes on consumption growth, and is the sum of health-induced changes in earnings, savings, transfers, and

[23] Estimating equation (6) via Probit instead of OLS yields a similar marginal effect of .398. The coefficient estimate on campaign remains stable when different controls are introduced. This lends credibility to the randomization process: in expectation treatment and control groups are similar across observed and unobserved characteristics and including covariates does not change the point estimate but increases precision.

OOP. This is analogous to theoretical equations (2) and (4). I estimate the reduced-form effect of health changes on consumption:[24]

$$(7) \qquad \Delta ln\left(\frac{C_{hv}}{n_{hv}}\right) = \gamma_v + \gamma_1 \Delta index_{hv} + X'_{hv}\gamma_2 + u_{hv}$$

which is a regression of the growth in log per capita (non-medical) consumption for household h in village v against the change in health ($\Delta index_{hv}$), a series of household demographic controls to capture secular trends in consumption (X_{hv}), and a random error (u_{hv}). I include cluster match fixed effects γ_v to control for changes in potential determinants of both income and health. Errors are clustered at the health cluster level.

The model is a fixed-effects specification and controls for time-invariant heterogeneity in omitted characteristics, such as risk preferences and health endowments. I discuss concerns about time-varying omitted variables and the endogeneity of health changes in Section 8.

If people are fully insured, changes in health will not affect consumption growth: $\gamma_2 = 0$. Results are displayed in Table 8 columns (1) and (2). Changes in the index have a significant and sizeable effect in the expected direction: moving from being able to perform all to unable to perform any activities lowers annualized PCE by 14%. This implies that net savings plus transfers are less than the lost labor earning and medical costs of illness. These results are smaller than those in Gertler and Gruber (2002): they find that a full health deterioration of the household head lowers annualized PCE by 20%. When I focus on health shocks to the household head in section 7.2, I find larger effects than in the main sample: a full health deterioration of the head lowers annualized PCE by 25%.

6.3 Consumption-Smoothing Benefit of HI

The previous section shows that households do not have full consumption insurance. Therefore, social insurance may provide a consumption smoothing benefit as opposed to simply crowding out informal insurance mechanisms. To test for the effect of health changes on consumption growth by HI status, I modify my change-in-health variable. I create a health deterioration index (health improvement index) that is equal to the index change if the change is negative (positive), and 0 otherwise. I include the health indices, their interactions with HI, and an HI dummy. I create separate health indices because adding an interaction term to equation (7) between the change in

[24] In order for there to be an effect of illness on consumption through imperfect insurance, there must be a sizable cost of illness. While I do not have data to estimate the effect of illness on labor earnings, I can estimate an OOP health expenditure equation by replacing the dependent variable in equation (7) with OOP health expenditures. There are significant effects in the expected direction: moving from completely able to unable to perform activities increases yearly medical spending by 5921 pesos. This is equivalent to twice the average yearly OOP.

health and HI could be misleading in that it restricts the sign and magnitude of the interaction coefficient to be the same regardless of the direction of health changes. There is no theoretical reason for this to be the case, as shown in Section 3. The excluded group is that with no health changes and no HI. I instrument for HI with treatment group, and the interaction terms with interactions between the health indices and treatment group. I estimate the following equation:

$$\Delta ln\left(\frac{C_{hv}}{n_{hv}}\right) = \beta_v + \beta_0 HI_{hv} + \beta_1 \Delta deteriorate_{hv} + \beta_2 \Delta improve_{hv} + \beta_3 \Delta deteriorate_{hv} * HI_{hv} + \beta_4 \Delta improve_{hv} * HI_{hv} + X'_{hv}\beta_5 + e_{hv} \quad (8)$$

which is a 2SLS regression of the growth in log per capita non-medical consumption for household h in village v against the change in health ($\Delta deteriorate_{hv}$ and $\Delta improve_{hv}$), HI status (HI_{hv}), the interaction of the change in health and HI status ($\Delta deterioration_{hv} * HI_{hv}$ and $\Delta improve_{hv} * HI_{hv}$), a series of household demographic controls (X_{hv}), cluster match fixed effects (β_v) and an error (e_{hv}). As in the previous regression, I include cluster match fixed effects β_v to control for changes in potential determinants of both income and health and standard errors are clustered at the health cluster level.

The coefficients on the interaction terms, β_3 and β_4, capture the differential impact of illness and improvements, respectively, among individuals with and without HI. From the previous section, I expect $\beta_1 < 0$ and $\beta_2 > 0$: health deteriorations (improvements) cause a decrease (increase) in consumption. The extent to which HI offsets illness-induced changes in consumption depends on how large medical expenditures are relative to the indirect effects of HI on informal mechanisms and the effect of health on labor earnings. This can be inferred from how much β_3 offsets β_1. The precautionary savings effect due to insurance for the households of those individuals who exhibit no change in health can be inferred from the coefficient on HI and I expect $\beta_0 > 0$.

Results are presented in columns (3) and (4) of Table 8.[25] For individuals without HI, complete health deteriorations decrease annualized household PCE by 30%. HI offsets this illness-induced decline in consumption by about 50%: for individuals with HI, complete health deteriorations decrease annualized household PCE by 14%. The estimates imply that the additional resources from informal insurance mechanisms used by the uninsured are not substantial enough to cover medical expenditures and next period's precautionary savings. Improvements increase consumption by .4% and 12% for individuals without and with HI, respectively. The increase is through a combination of decreased medical expenditures and increased earnings. The difference

[25] Table 10 shows the first stage regressions.

between the estimates imply that a combination of less transfers and higher next period precautionary savings overwhelm the health-induced gain in resources for the uninsured. For example, the uninsured may allocate more of their health-improvement induced earnings increases toward precautionary savings compared to the insured. In fact, when I limit my sample to individuals who are employed at baseline in section 7.3, there is a larger differential effect of HI for health improvements. This lends credibility to my hypothesis because the effect of health on labor earnings is likely larger for the employed. For the households of individuals who exhibit no change in health, annualized household PCE increases by 4.6%, likely due to a decrease in current income allocated toward precautionary savings. This hypothesis is consistent with the existing literature on consumption smoothing in Mexico, in that households decrease savings when they have access to formal insurance mechanisms (Ruiz, 2011).

6.4 Health Care Financing

The data does not include information on income, earnings, savings, or transfers and thus precludes analysis of the direct mechanisms through which HI moderates the effect of health on consumption- via labor earnings, savings and inter-household transfers. However, the survey does contain questions regarding the various methods the household uses to finance the previous years' medical expenditures: whether from current income, by drawing from savings, by selling assets or borrowing. While I am unable to examine the *change* in household savings or transfers due to health changes differentially by insurance status, I can ascertain if health care financing methods vary across insurance status. For example, whether uninsured households are more likely to borrow to finance their medical expenditures compared to the insured.

I estimate the effect of insurance status in the follow-up period on the OOP financing method y_{hv} for household h in village v with the following instrumental variables regression:

$$(9) \qquad y_{hv} = \pi_v + \pi_1 HI_{hv} + X'_{hv}\pi_2 + u_{hv}$$

where y_{hv} is an indicator for whether the household financed medical expenditures through current income, by drawing from savings, by selling assets and by borrowing in the follow-up period. I instrument for HI using the campaign as an instrumental variable and the first-stage is shown in Table 7. The vector of household controls X_h includes: household size, urban/rural location, whether the household is indigenous, and Oportunidades enrollment. I include cluster match fixed effects and cluster my standard errors at the health cluster level.

The coefficient π_1 on HI gives the average difference in means between the insured and uninsured. As shown in Table 9, insurance status has a significant effect on the likelihood of drawing

from savings or selling assets to finance medical expenditures. Insured households are 4.6 percentage points less likely to draw from savings than uninsured households. Results are similar for financing by selling assets: insured households are 4.2 percentage points less likely to sell assets. Finally, while the effects of insurance on financing from current income or borrowing are negative, they are imprecisely estimated.

7. Heterogeneous Effects

In this section, I consider the consumption-smoothing benefit of HI for three groups: Oportunidades households, households for whom the randomly selected individual is the household head, and households for whom the randomly selected individual work in the labor market at baseline.

7.1 Oportunidades Households

Oportunidades is a conditional cash transfer program for families with school-age children. The effect of HI on Oportunidades families is of particular interest because these households are in the lowest two income deciles, automatically enrolled in SP, and receive basic preventative health care and cash transfers through Oportunidades. The differential effect of HI on health-induced changes in consumption could be either smaller or larger for these households. Because they are in the lowest 2 income deciles, OOP are more likely to be catastrophic in treating a given illness and I expect HI to provide more consumption-smoothing benefits at the lower end of the income distribution. However, all Oportunidades households receive free basic preventative health care and thus may have lower OOP compared to the entire sample. In addition, households receive a cash transfer equal to about 20% of monthly household income. This cash transfer may cushion the decrease in consumption due to income shocks, such as health-induced changes in labor income.[26]

Oportunidades households comprise 47% of the sample. Descriptive statistics of individuals in Oportunidades households are shown in Table 11. Individuals are slightly less educated, less likely to be employed, and less likely to have HI at baseline compared to the entire sample.

I estimate equation (7), limiting observations to Oportunidades households. The results are presented in Table 12. Columns (1) and (2) indicate that a full health deterioration has a slightly smaller effect on PCE than in the full sample: moving from being able to perform all to unable to perform any activities lowers monthly food expenditure by 17.8% and there is no statistically

[26] Skoufias (2002) finds that the early, rural implementation of Oportunidades, PROGRESA, does enable households to better smooth consumption over income fluctuations.

significant effect on annualized PCE. This may be because the cash transfer received through Oportunidades better insulates these households from adverse events, such as health shocks.

The differential effect of HI for Oportunidades households is presented in columns (3) and (4). For the households of individuals who exhibit no change in health, annualized household PCE increases by 8.1%, likely due to a decrease in current income allocated toward precautionary savings. This is nearly twice as large as for the entire sample, and consistent with the findings of Ruiz (2011): the introduction of formal credit institutions causes a decrease in household savings, with the poorest households exhibiting the highest decline in saving rates. For individuals without HI, complete health deteriorations lower annualized household PCE by 28%. This is a slightly smaller effect than for non-Oportunidades households because Oportunidades households receive a cash transfer which may buffer these families from income fluctuations. HI offsets this illness-induced decline in consumption by 75%: for individuals with HI, complete health deteriorations decrease annualized household PCE by 7%.

The Oportunidades program combined with HI provides substantial protection for this particularly vulnerable segment of the population from negative health shocks. It is likely that HI assuages the effect of increased total medical expenditures due to negative health shocks, while the cash transfer protects the household against income loss due to illness.[27] This hypothesis is consistent with previous findings on the consumption smoothing benefit of Oportunidades (Skoufias, 2002). The differential effects of HI for health improvements of individuals in Oportunidades households are larger than in the entire sample: a full health improvement increases annualized PCE by 2% for the uninsured and 13% for the insured. When health improvements induce an increase in labor earnings and decrease in OOP, it may be that Oportunidades households allocate fewer resources toward precautionary savings because of the buffer provided by the cash transfer. The differential effect of HI across health improvements is the same for Oportunidades households (11%) as in the general sample.

7.2 Health Shocks to the Household Head

The data contain detailed individual health data from one randomly selected adult in each household. In this section, I limit my results to households for whom the randomly selected individual is the household head, forty-seven percent of the sample. Descriptive statistics for this group are shown in Table 11. Sixty-two percent of household heads are male, compared to 36 % in

[27] Because individuals in Oportunidades are slightly less likely to be employed at baseline compared to the entire sample (47% and 51% employment rates, respectively), the effect of health on consumption through labor earnings is likely not as large, absent any social programs, as in the entire sample.

the entire sample, and they are five years older than individuals in the main sample, on average. Household heads have more education than individuals in the entire sample and 75% are employed. The effect of a health shock to the household head on household consumption may be larger because these individuals likely contribute more to household income than other individuals, as evidenced by their higher employment rates. I expect the differential effect of HI for this group to be no larger than that found in the entire sample because HI moderates the effect of illness on consumption through its effect on medical expenditures and does not insure against the loss in labor earnings associated with adverse health events. For a group with a high employment rate, the loss in labor earnings is likely substantial. I focus specifically on the heterogeneous effects for the employed in the next section.

I estimate equation (7), limiting observations to households where the individual whose health changes I measure are the household head. The results are presented in Table 13. Columns (1) and (2) indicate that a full health deterioration of the head has a substantially larger effect on PCE than in the full sample: moving from being able to perform all to unable to perform any activities lowers monthly food expenditure by 25.7% and annualized PCE by 25.4%. These results are slightly larger than those found by Gertler and Gruber (2002).

The consumption-smoothing effects of HI are presented in columns (3) and (4). For the households of individuals who exhibit no change in health, annualized household PCE increases by 5.2%. This is likely due to a decrease in current income allocated toward precautionary savings and is comparable to results for the entire sample. For the uninsured, complete health deteriorations decrease monthly food expenditure by 30% and lower annualized household PCE by 32%. HI offsets this illness-induced decline in consumption by less than half: complete health deteriorations decrease annualized household PCE by 18% for the insured. It is not surprising that HI has a smaller consumption smoothing benefit against health shocks to the household head: HI does not insure against the loss in labor income associated with adverse health events, and thus when the primary earner is ill, total household income less medical expenditures decreases substantially. For households whose head experiences a full health improvement, annualized PCE increases by 2% and 21% for the uninsured and insured. The latter estimate is considerably larger than that in the entire sample (12%) and is consistent with the hypothesis that uninsured and insured households differ in the extent to which they allocate health improvement-induced increases in labor earnings toward precautionary savings. Because the household head is often the primary income earner, a full

health improvement is associated with higher consumption growth than that found in the entire sample. My findings in the next section corroborate this hypothesis.

7.3 Health Shocks to the Employed

In this section, I consider the consumption-smoothing benefits of HI for households whose randomly selected individual is employed at baseline, about 51% of the sample. Illness induces both high OOP and a loss in labor income, and HI only assuages the effect of illness on consumption through its effect on OOP. Therefore, HI may not provide as large of a consumption-smoothing benefit when I focus exclusively on the employed. The differential effect for the employed depends on how much intra-household reallocation of market and household work occurs in response to health shocks. For example, if the employed individual experiences a deterioration of their health, opts out of the labor force, and instead spends their time in household production, another household member may increase their labor market participation to fully offset the losses in labor income incurred by the sick household member. In this case, HI may provide as much consumption-smoothing benefit as for an individual who was not employed at baseline. However, estimates on the heterogeneous effects for household heads in the previous section suggest that this is not the case.

I restrict the sample to individuals who are employed at baseline. Table 11 shows the descriptive statistics for these individuals. On average they are 43 years old and 63% are male, compared to 36% male in the entire sample. I estimate equation (7), limiting observations to households where the individual whose health changes I measure is employed at baseline. Results in Table 14 columns (1) and (2) indicate that a full health deterioration of the employed has a slightly larger effect on PCE than in the full sample: complete health deteriorations lower monthly food expenditure by 18.7% and annualized PCE by 16.6%. These results are considerably smaller than if the individual is the household head, likely due to the fact that the household head is often the primary earner.

The differential effects of HI are presented in columns (3) and (4). For the households of individuals who exhibit no change in health, annualized household PCE increases by 4.6%, comparable to results for the entire sample. For the uninsured, complete health deteriorations decrease monthly food expenditure by 27% and lower annualized household PCE by 28%. Similar to the consumption smoothing effect for household heads, HI offsets this illness-induced decline in consumption by less than half: complete health deteriorations decrease annualized household PCE by 15.2% for the insured. For households whose head experiences a full health improvement,

annualized PCE increases by 2% and 25% for the uninsured and insured. The differential effects for the household head and employed samples are very similar and this underscores the importance of the uninsured economic cost associated with health, labor earnings. HI does not fully offset the economic costs associated with illness and my estimates suggest that public disability insurance may improve welfare by insuring households against the labor earnings losses associated with illness.

8. Validity Concerns

My theoretical and empirical work requires relies on the assumption that health insurance does not directly affect health or health care utilization. Given that baseline health is not different across treatment and control groups, health insurance would have to significantly affect follow-up health status, as measured by the health index, for this to be a serious empirical concern. In Table 15, I provide instrumental variable estimates of the effect of HI on health and health care utilization as measured by hospitalizations and outpatient care. Consistent with King et al. (2007), I do not find an effect of HI on health or health care utilization. This is likely due to the short evaluation period. One might expect to find an effect on health and utilization over a longer time horizon, in which case my empirical strategy would be inappropriate.

I remain concerned about omitted variables, endogeneity of health changes, and reverse causality. Changes in health reflect choices of endogenous health inputs and exogenous unobserved health shocks. The endogenous inputs reflect income and price shocks and therefore the estimating equation needs good proxies for these. Weather is potentially confounding because rainfall or temperature shocks may be correlated with health and income. I cannot use community FE to control for income, price and weather shocks because the treatment is at the community level. Instead, I control for the health cluster match. I assume that 2 communities that looked similar at baseline trend in the same way in terms of potential determinants of both income and health over the evaluation.

If individuals who experience health shocks are different in an unobserved way that is correlated with ability to smooth consumption, my estimates will be bias. For example, preference heterogeneity that leads the people who undertake preventive health care measures and face a lower illness probability to save more. The majority of the literature does not address this issue.[28] Following Finkelstein and McGarry (2006), I examine the correlation between health changes and

[28] Papers that ignore this include: Gertler and Gruber (2002), Gruber et al. (2009), Townsend (1994), Kochar (1995), Dercon and Krishnan (2000), De Weerdt and Dercon (2006), Finkelstein et al. (2008), and Foster (1995). The only paper that studies negative health shocks and accounts for endogeneity is Mohanan (2011). His findings on consumption smoothing over small, health changes are consistent with the previous literature that does not correct for selection.

preventive behaviors, such as receiving a flu shot. Results are presented in Table 16. I find that there is a significant relationship between health changes in preventive behaviors at baseline. Health improvements are positively correlated with receiving a flu vaccine and eye exam in the baseline year. Health improvements are negatively and insignificantly correlated with being a smoker at baseline. All correlations are relatively small in magnitude. However, these results imply that there is a potential issue related to adverse selection. If individuals who experience adverse health events are less likely to undertake preventive health measures, and are also less likely to save for precautionary reasons, then we may observe a larger consumption-smoothing benefit of HI than for individuals who are more cautious and do not experience health deteriorations. Alternatively, if individuals who do not undertake preventive health measures are more likely to save because they know they are at an increased risk for adverse health events, we may observe a smaller consumption smoothing benefit of HI. Without data on savings or other informal insurance mechanisms it is difficult to determine the extent of potential adverse selection and the direction of bias is ambiguous.

Reverse causality is an issue if changes in income affect health, for example, job loss that results in health deterioration due to depression. If my results reflect the effect of labor supply on health, this mechanism would need to operate more strongly the larger the shock to labor earnings: the effect on health from reduced earnings would be bigger for high-wage versus low-wage individuals. While unlikely, I cannot test for this because I do not observe labor earnings or wages.

I ignore the possibility of state dependence in my theoretical and empirical work: I assume that the marginal utility of consumption evaluated at the same level of consumption is the same regardless of health state. Allowing for state dependence adds a term to consumption growth Δc that represents the change in consumption purely due to state dependence:

$$\Delta c = c_2^* - c_1^* = t - (1-k)w - (1-b)m + (2+r){s_1}^*(w,k,p,b,m,t) - {s_{2s}}^*(w,k,p,b,m,t)$$

This additional term adds a scaling effect: the empirical results will capture changes in consumption due to health *and* changes in consumption due to changes in the marginal utility of consumption due to changes in health. While state dependence changes the interpretation of the magnitude of my estimates, it does not change the sign on the interaction coefficients because the marginal utility of consumption in different health states wouldn't vary by HI. The magnitude of the interaction coefficient would provide the relative effect of HI. For example, consider the extreme case with negative state dependence where the 14% decline in PCE for the insured is due to state dependence.

This would imply households with HI are fully insured against health shocks. The (negative) consumption growth for the uninsured could then be decomposed into two parts: a 14% decrease in consumption due state dependence, and a 16% decrease in consumption due to lack of full insurance.

9. Conclusion

Major illness exposes families to substantial financial risks and it is not surprising that uninsured families are unusually vulnerable to adverse health shocks. For this reason, many countries are considering social insurance to help insure families against the economic costs of illness. Using experimental data from one of the largest expansions of public health insurance in the world, I assess the extent to which insurance cushions health-induced changes in non-medical consumption in rural Mexico. I find that HI offsets illness-induced declines in consumption by over 50%. In addition, there are substantially larger effects for Oportunidades households, for whom HI offsets illness-induced declines in consumption by 75%.

The results of this paper have important policy implications. In the short term, the social insurance program did not affect several outcomes that were important to policymakers, such as health care utilization or health. However, the program did significantly reduce the incidence of catastrophic medical expenditure and provide a consumption-smoothing benefit to families who faced serious health shocks. HI does not fully offset the economic costs associated with illness because it does not insure against the labor earnings losses: HI offsets health deterioration-induced declines in consumption by less than 50% for employed individuals. These estimates suggest that public disability insurance may provide further benefits by insuring households against the labor earnings losses associated with illness. The Mexican experience with social insurance may inform the construction of social insurance programs in other countries.

References

Agartan, T., 2005. Health Sector Reform in Turkey: Old Policies New Politics, mimeo.

Javier Arias, Oliver Azuara, Pedro Bernal, James J. Heckman, and Cajeme Villar. Policies to promote growth and economic efficiency in Mexico. Working Paper 16554, National Bureau of Economic Research, November 2010.

Azuara, Oliver and Marinescu, Ioanna (2011). "Informality and the Expansion of Social Protection Programs: The Case of Mexico." Unpublished manuscript, University of Chicago.

Barofsky, Jeremy. 2011. "Estimating the impact of health insurance in developing nations: evidence from Mexico's Seguro Popular". Working Paper, Harvard University

Rodrigo Barros. Wealthier but not much healthier: Effects of a health insurance program for the poor in Mexico. Working paper, Stanford University, 2011.

Deaton, A., 1992a. Household saving in LDCs: credit, markets, insurance and welfare. *Scandinavian Journal of Economics* 94 (2), 253–273.

Deaton, A., 1992b. Understanding Consumption. Oxford University Press, New York.

Deaton, A., 1997. The Analysis of Household Surveys, The World Bank, Baltimore: John Hopkins.

Dercon, S., Krishnan, P., 2000. In sickness and in health: risk-sharing in rural Ethiopia. *Journal of Political Economy* 108 (4), 688– 727.

De Weerdt, Joachim & Dercon, Stefan, 2006. Risk-sharing networks and insurance against illness. *Journal of Development Economics* 81(2), 337-356.

Finkelstein, Amy, Erzo Luttmer, and Matthew Notowidigdo, 2008. What Good is Wealth without Health? The Effect of Health on the Marginal Utility of Consumption. NBER Working Paper No. 14089.

Finkelstein, Amy, and Kathleen McGarry, 2006. Multiple Dimensions of Private Information: Evidence from the Long-Term Care Insurance Market. *American Economic Review* 96(4), 938–58.

Foster, Andrew, 1995. Prices, Credit Markets and Child Growth in Low-Income Rural Areas. *The Economic Journal* 105(430), 551-70.

Julio Frenk. Bridging the divide: global lessons from evidence-based health policy in Mexico. Lancet, 368(9539):954-961, Sep 2006.

Julio Frenk, Eduardo Gonzlez-Pier, Octavio Gmez-Dants, Miguel A Lezana, and Felicia Marie Knaul. Comprehensive reform to improve health system performance in Mexico. Lancet, 368(9546):1524-1534, Oct 2006.

Gakido, Emmanuela, Rafael Lozano, Eduardo González-Pier, Jesse Abbott-Klafter, Jeremy T. Barofsky, Chloe Bryson-Cahn, Dennis M. Feehan, Dianna K. Lee, Hector Hernández-Llamas, Christopher and J.L. Murray (2006). "Assessing the Effect of the 2001-06 Mexican Health Reform: An Interim Report Card," *Lancet, 368*, 1920-1935.

Gertler, Paul, and Jonathan Gruber, 2002. Insuring Consumption Against Illness. *American Economic Review* 92(1), 51–70.

Gertler, P., D. I. Levine, and E. Moretti, 2009. Do microfinance programs help families insure consumption against illness? *Health Economics* 18, 257-273.

Gruber, Jonathan, 2000. Cash welfare as a consumption smoothing mechanism for single Mothers. *Journal of Public Economics* 75(2), 157-82.

Gruber, J., 2008. Covering the Uninsured in the United States," *Journal of Economic Literature* 46(3), 571-606.

Homedes, Nuria, and Antonio Ugalde (2009). "Twenty-Five Years of Convoluted Health Reforms in Mexico," *PLoS Medicine, 6*, e1000124

Kazianga, Harounan & Udry, Christopher, 2006. Consumption smoothing? Livestock, insurance and drought in rural Burkina Faso. *Journal of Development Economics* 79(2), 413-446.

King, Gary, Emmanuela Gakidou, Nirmala Ravishankar, Ryan T. Moore, Jason Lakin, Manett Vargas, Martha María Téllez-Rojo, Juan Eugenio Hernández-Ávila, Mauricio Hernández-Ávila, and Hector Hernández Llamas. 2007. A 'Politically Robust' Experimental Design for Public Policy Evaluation, with Application to the Mexican Universal Health Insurance Program. *Journal of Policy Analysis and Management* 26, 479-506.

King, Gary, Emmanuela Gakidou, Kosuke Imai, Jason Lakin, Ryan T. Moore, Clayton Nall, Nirmala Ravishankar, Manett Vargas, Martha María Téllez-Rojo, Juan Eugenio Hernández-Ávila, Mauricio Hernández-Ávila, and Hector Hernández Llamas. 2009. Public Policy for the Poor? A Randomised Assessment of the Mexican Universal Health Insurance Programme. *Lancet* 373, 1447-1454.

Felicia Marie Knaul, Hctor Arreola-Ornelas, Oscar Mndez-Carniado, Chloe Bryson- Cahn, Jeremy Barofsky, Rachel Maguire, Martha Miranda, and Sergio Sesma. Evidence is good for your health system: policy reform to remedy catastrophic and impoverishing health spending in Mexico. The Lancet, 368(9549):1828-1841, 2006.

Knox, Melissa A. (2008). "Health Insurance for All: An Evaluation of Mexico's Seguro Popular Program." Unpublished Manuscript. University of California, Berkeley.

Kochar, A., 1995. Explaining Household Vulnerability to Idiosyncratic Income Shocks. *American Economic Review 85*(2), 159– 64.

Jason Lakin. The end of insurance? Mexico's Seguro Popular, 2001 2007. Journal of Health Politics, Policy and Law, 35(3):313{352, June 2010.

Mohanan, M., 2011. Causal Effects of Health Shocks on Consumption and Debt: Quasi-Experimental Evidence from Bus Accident Injuries. Forthcoming *Review of Economics and Statistics.*

Rosenzweig, M.R., Wolpin, K.I., 1993. Credit market constraints, consumption smoothing, and the accumulation of durable production assets in low-income countries: investments in bullocks in India. *Journal of Political Economy* 101 (2), 223– 244.

Spenkuch, Jorg L., 2011, "Moral Hazard and Selection Among the Poor: Evidence from a Randomized Experiment", forthcoming *Journal of Health Economics*.

Sandra G. Sosa-Rubi, Omar Galarraga, and Jerey E. Harris. Heterogeneous impact of the program on the utilization of obstetrical services in Mexico, 2001-2006: A multinomial probit model with a discrete endogenous variable. Journal of Health Economics, 28(1):20-34, 2009.

Thornton, Rebecca, Laurel Hatt, Erica Field, Mursaleena Islam, Freddy Solís, Martha Azuzena González Moncada. 2010. Social Security Health Insurance for the Informal Sector in Nicaragua: A Randomized Evaluation. *Health Economics* 19, 181.206.

Townsend, R.M., 1994. Risk and Insurance in Village India. *Econometrica* 62 (3), 539–591.

Trujillo et al., 2005. The Impact of Subsidized Health Insurance for the Poor: Evaluating the Colombian Experience Using Propensity Score Matching. *International Journal of Health Care Finance and Economics* 5, 211-239.

Wagstaff, A., 2007. Health Insurance for the Poor: Initial Impacts of Vietnam's Health Care Fund for the Poor. World Bank Policy Research Working Paper Series 4134.

Wagstaff et al., 2007. Extending Health Insurance to the Rural Population: an Impact Evaluation of China's New Cooperative Medical Scheme. World Bank Policy Research Working Paper Series 4150.

Table 1: Descriptive Statistics at Baseline

	Total Sample				Control	Treatment	Difference in Means (C-T)
	Mean	Std. Dev.	Min	Max	Mean	Mean	
Individual Characteristics							
Age	42.83	16.62	18.00	98.00	42.75	42.91	-0.16
Sex (male)	0.36	0.48	0.00	1.00	0.36	0.35	0.01**
Educ: primary	0.32	0.47	0.00	1.00	0.32	0.32	0.00
Educ: secondary	0.25	0.43	0.00	1.00	0.25	0.24	0.01*
Educ: vocational	0.24	0.43	0.00	1.00	0.23	0.25	-0.02**
Married	0.58	0.49	0.00	1.00	0.58	0.57	0.01
Dependents	0.56	0.50	0.00	1.00	0.56	0.56	0.00
Employed	0.51	0.50	0.00	1.00	0.50	0.51	-0.01*
HI Status	0.21	0.41	0.00	1.00	0.18	0.25	-0.08**
Household Characteristics							
HH Size	4.19	2.06	1.00	19.00	4.13	4.24	-0.11**
Rural	0.91	0.28	0.00	1.00	0.92	0.90	0.02**
Indigenous	0.06	0.24	0.00	1.00	0.06	0.07	-0.01*
Oportunidades	0.47	0.50	0.00	1.00	0.47	0.48	-0.01*
HH Consumption & Assets							
HH PCE Food (1 mth)	361	449	0.00	15500	358	364	-6.85
HH PCE Essential Items (1 mth)	550	601	0.00	20366	548	552	-4.13
HH PCE All Items (1 mth)	809	1373	0.00	99857	800	818	-17.55
Annualized HH PCE	9092	10896	0.00	368320	8964	9220	-255.89
Asset Index	0.46	0.19	0.00	1.00	0.45	0.47	-0.02**
HH Health Expenditures							
HH Health Exp. (1 mth)	157	553	0.00	10,000	160	155	5.82
HH Health Exp. (3 mth)	677	1,645	0.00	37,000	677	676	1.24
HH Health Exp. (12 mth)	3,535	9,375	0.00	903,520	3,565	3,505	59.93
Proportion of Annual Spending on Health	0.09	0.14	0.00	1.00	0.09	0.09	0.00
Catastrophic Health Expenditures	0.09	0.28	0.00	1.00	0.09	0.09	0.00
Health Care Utilization							
No. HH Hospitalizations	0.21	0.59	0.00	9.00	0.21	0.20	0.00
Outpatient care (12 mth)	0.66	0.47	0.00	1.00	0.66	0.65	0.01
Hospitalization (12 mth)	0.07	0.26	0.00	1.00	0.07	0.07	0.00

* $p<0.1$; ** $p<0.05$; *** $p<0.01$

Table 2: Health Variables at Baseline

	Total Sample				Control	Treatment	Difference in Means (C-T)
	Mean	Std. Dev.	Min	Max	Mean	Mean	
Smoker	0.10	0.31	0.00	1.00	0.11	0.10	0.01*
Chronic disease	0.16	0.36	0.00	1.00	0.16	0.16	0.00
BP: pre-hypertension	0.38	0.49	0.00	1.00	0.38	0.38	0.00
BP: stage 1 hypertension	0.12	0.32	0.00	1.00	0.12	0.12	0.00
BP: stage 2 hypertension	0.06	0.23	0.00	1.00	0.05	0.06	0.00
BMI: Severely underweight	0.007	0.083	0.00	1.00	0.007	0.006	0.00
BMI: Underweight	0.016	0.124	0.00	1.00	0.015	0.015	0.00
BMI: Pre-obese	0.368	0.482	0.00	1.00	0.360	0.368	-0.01
BMI: Obese	0.173	0.378	0.00	1.00	0.176	0.168	0.01

Blood pressure and BMI categorized according to World Health Organization classifications. Chronic disease includes: heart disease, hypertension and diabetes. * p<0.1; ** p<0.05; *** p<0.01

Table 3: Self-Reported Health at Baseline

	Very Good	**Good**	**Moderate**	**Bad**	**Very Bad**
Rating of Overall Health	2,165	11,860	9,572	1,271	88
Activity	**None**	**Mild**	**Moderate**	**Severe**	**Can't do**
difficulty in moving around*	18,432	3,676	1,978	765	105
difficulty in vigorous activity (running 3 km or cycling) *	14,696	3,840	2,706	1,950	1,764
difficulty with self-care (washing, dressing) *	23,135	1,190	458	142	31
difficulty in maintaining appearance (grooming) *	12,951	7,002	3,457	1,406	140
difficulty in feeding yourself (cooking, using utensils)*	13,228	7,039	3,260	1,304	125
difficulty concentrating or remembering things	21,780	2,114	757	243	62
difficulty of learning new tasks	22,158	1,888	661	199	50
difficulty with relationships & community participation	18,722	3,849	1,666	572	147
difficulty in getting along with others	16,025	5,172	2,624	1,015	120
difficulty in performing work or other regular activities*	14,993	6,119	2,755	952	137
Pain & Discomfort	**None**	**Mild**	**Moderate**	**Severe**	**Extreme**
rating of pain & bodily aches	14,765	6,045	2,812	1,207	127
rating of soreness & discomfort	16,238	5,835	2,157	673	53
difficulty in daily life due to pain	17,080	5,145	1,979	668	84
difficulty with sleep	16,380	5,420	2,109	874	173

Table 4: Summary Statistics of Baseline & Changes in Health Index

	Baseline	Change	Min	Max
Health Index	0.853	0.063	-1	0.917
	(0.1646)	(0.1641)		
Some limitation	0.685	-0.181	-1	1
Any deterioration (=1)		0.220	0	1
Mean deterioration conditional		-0.231	-1	-0.042
on some deterioration		(0.1563)		
Any improvement (=1)		0.532	0	1
Mean improvement conditional		0.273	0.042	0.917
on some improvement		(0.1487)		

Table 5: Health Insurance Status by Type

	Baseline	Follow-up
Health Insurance	0.215	0.428
Type Conditional on HI		
Obligatory	0.636	0.312
Voluntary, non-SP	0.056	0.062
Voluntary, SP	0.264	0.614
Obligatory & Voluntary	0.036	0.01
HI Type Missing	0.008	0.001

Table 6: Change in HI by Treatment Group

	Control	Treatment
No HI- No HI	0.73	0.34
No HI- HI	0.10	0.40
HI- HI	0.13	0.22
HI- No HI	0.04	0.03
Total	12,382	12,574

Table 7: Effect of Encouragement Campaign on Health Insurance Status

Dependent Variable is HI	(1)	(2)	(3)
Campaign	0.391	0.389	0.386
	(68.60)***	(68.35)***	(68.24)***
Household size		-0.002	-0.003
		(1.16)	(1.99)**
Rural		-0.107	-0.107
		(10.45)***	(10.44)***
Indigenous		0.104	0.110
		(8.63)***	(9.09)***
Oportunidades		0.117	0.127
		(19.46)***	(20.68)***
Age			0.005
			(4.85)***
Age Squared			-0.000
			(2.68)***
Sex (male)			-0.051
			(6.79)***
Educ: primary			0.067
			(7.45)***
Educ: secondary			0.083
			(8.37)***
Educ: college/vocational			0.150
			(14.27)***
Married			0.081
			(13.29)***
Dependents			0.017
			(2.63)***
Employed			0.020
			(2.77)***
Community Match FE		x	x
R^2	0.16	0.18	0.19
N	24,956	24,956	24,956

* $p<0.1$; ** $p<0.05$; *** $p<0.01$

Table 8: Effect of Health Changes on Consumption Growth

	(1) Change in Log PCE Food (1m)	(2) Change in Log PCE (12m)	(3) Change in Log PCE Food (1m)	(4) Change in Log PCE (12m)
HI			0.070	0.046
			(2.09)**	(3.00)***
Change in index	0.186	0.140		
	(3.89)**	(3.13)**		
Index decreases			-0.296	-0.306
			(2.01)**	(2.46)**
Interact: Index decreases, HI			0.180	0.164
			(3.67)***	(2.81)***
Index increases			0.007	0.004
			(0.07)	(2.28)**
Interact: Index increases, HI			0.141	0.119
			(1.90)*	(2.76)***
Household Size	0.001	-0.022	0.001	-0.023
	(.55)	(11.11)***	(0.40)	(11.19)***
Urban	0.022	0.033	0.03	0.039
	(1.35)	(2.27)**	(1.70)*	(2.67)**
Indigenous	0.098	0.064	0.091	0.058
	(4.91)***	(3.80)***	(4.49)***	(3.42)***
Oportunidades	0.057	0.021	0.048	0.012
	(5.72)***	(2.49)**	(4.54)***	(1.37)
Community Match FE	x	x	x	x
R^2	0.04	0.05	0.00	0.01
N	24,956	24,956	24,956	24,956

* $p<0.1$; ** $p<0.05$; *** $p<0.01$

Table 9: IV Estimates of the Effect of HI on Health Care Financing Methods

	Income	Savings	Selling Assets	Borrowing
HI	-0.029	-0.046	-0.042	-0.024
	(0.51)	(2.16)**	(2.41)**	(0.85)
Household Size	0.015	0.000	0.001	0.003
	(9.79)***	(0.52)	(0.61)	(2.11)**
Urban	0.324	0.050	0.023	0.012
	(14.16)***	(2.30)**	(0.96)	(0.34)
Indigenous	-0.073	-0.028	0.003	-0.047
	(4.80)***	(2.07)**	(0.25)	(2.77)***
Oportunidades	-0.042	-0.011	0.011	-0.010
	(3.58)***	(1.91)*	(2.40)**	(1.40)
Community Match FE	x	x	x	x
R^2	0.06	0.03	0.03	0.04
N	24,956	24,956	24,956	24,956
Average	0.453	0.098	0.072	0.189

* $p<0.1$; ** $p<0.05$; *** $p<0.01$

Table 10: First Stage Regressions

	HI	Δdeteriorate*HI	Δimprove*HI
Campaign	0.37	0.00	0.00
	(46.89)**	(2.66)*	(3.08)*
Δdeteriorate*campaign	-0.08	0.42	-0.02
	(0.96)	(66.81)**	(1.41)
Δimprove*campaign	0.21	-0.01	0.46
	(0.83)	(1.45)	(63.89)**
Δdeteriorate	0.08	0.20	0.01
	(1.11)	(46.03)**	(1.06)
Δimprove	-0.09	0.01	0.20
	(1.23)	(1.65)	(28.82)**
HH Size	0.000	0.000	0.000
	(1.23)	(2.77)*	(1.36)
Rural	-0.11	0.000	-0.01
	(10.42)**	(6.27)**	(4.15)**
Indigenous	0.11	0.000	0.01
	(8.73)**	(2.80)*	(3.59)*
Oportunidades	0.12	0.000	0.01
	(19.39)**	(6.30)**	(12.8)**
Community Match FE	x	x	x
R^2	0.18	0.48	0.46

Table 11: Descriptive Statistics at Baseline for Heterogeneous Groups

	Entire Sample	Oportunidades	Household Head	Employed
	Mean	Mean	Mean	Mean
Age	42.83	42.79	47.42	43.02
Sex (male)	0.36	0.33	0.62	0.63
Educ: primary	0.32	0.35	0.34	0.32
Educ: secondary	0.25	0.25	0.33	0.24
Educ: vocational	0.24	0.17	0.21	0.27
Married	0.58	0.61	0.53	0.55
Dependents	0.56	0.57	0.54	0.51
Employed	0.51	0.47	0.75	
HI Status	0.21	0.17	0.22	0.23
HH Size	4.19	4.71	3.67	4.12
Rural	0.91	0.97	0.91	0.90
Indigenous	0.06	0.09	0.06	0.06
Oportunidades	0.47		0.46	0.45
N	24,956	11,862	11,854	12,667

Table 12: Effect of Health Changes on Consumption Growth for Oportunidades Households

	(1) Change in Log PCE Food (1m)	(2) Change in Log PCE (12m)	(3) Change in Log PCE Food (1m)	(4) Change in Log PCE (12m)
HI			0.108	0.081
			(2.97)***	(2.66)**
Change in index	0.178	0.110		
	(2.96)**	(0.12)		
Index decreases			-0.19	-0.283
			(1.02)	(1.81)*
Interact: Index decreases, HI			0.130	0.214
			(2.45)**	(2.09)**
Index increases			0.019	0.021
			(0.15)	(1.66)*
Interact: Index increases, HI			0.10	0.11
			(1.58)	(1.92)*
Household Size	0.006	-0.017	0.005	-0.018
	(1.79)*	(6.39)***	(1.57)	(6.58)***
Urban	0.092	0.09	0.107	0.102
	(2.05)**	(2.39)**	(2.38)**	(2.71)**
Indigenous	0.094	0.076	0.078	0.064
	(3.92)***	(3.77)***	(3.22)***	(3.11)**
Community Match FE	x	x	x	x
R^2	0.03	0.04	0.01	0.01
N	11,862	11,862	11,862	11,862

* $p<0.1$; ** $p<0.05$; *** $p<0.01$

Table 13: Effect of Household Head Health Changes on Consumption Growth

	(1) Change in Log PCE Food (1m)	(2) Change in Log PCE (12m)	(3) Change in Log PCE Food (1m)	(4) Change in Log PCE (12m)
HI			0.072	0.052
			(2.43)**	(2.05)**
Change in index	0.257	0.254		
	(4.44)***	(2.10)**		
Index decreases			-0.305	-0.326
			(2.44)**	(2.90)***
Interact: Index decreases, HI			0.125	0.142
			(2.22)**	(2.19)**
Index increases			0.071	0.020
			(0.49)	(1.68)*
Interact: Index increases, HI			0.130	0.194
			(2.32)**	(2.57)**
Household Size	0.005	-0.022	0.005	-0.018
	(1.28)	(7.15)***	(1.57)	(6.58)***
Urban	0.029	0.031	0.107	0.102
	(1.22)	(1.51)	(2.38)**	(2.71)**
Indigenous	0.095	0.044	0.078	0.064
	(3.37)***	(1.83)*	(3.22)***	(3.11)**
Community Match FE	x	x	x	x
R^2	0.03	0.04	0.01	0.02
N	11,854	11,854	11,854	11,854

* $p<0.1$; ** $p<0.05$; *** $p<0.01$

Table 14: Effect of Health Changes of the Employed on Consumption Growth

	(1) Change in Log PCE Food (1m)	(2) Change in Log PCE (12m)	(3) Change in Log PCE Food (1m)	(4) Change in Log PCE (12m)
HI			0.067	0.046
			(3.97)***	(3.01)***
Change in index	0.187	0.166		
	(2.64)***	(2.11)**		
Index decreases			-0.271	-0.276
			(2.72)**	(2.53)**
Interact: Index decreases, HI			0.114	0.124
			(2.50)**	(2.62)**
Index increases			0.039	0.025
			(0.25)	(1.81)*
Interact: Index increases, HI			0.135	0.227
			(1.65)*	(2.33)**
Household Size	0.004	-0.020	0.004	-0.020
	(1.35)	(7.19)***	(1.26)	(7.14)***
Urban	0.034	0.038	0.038	0.037
	(1.51)	(2.02)**	(1.69)*	(1.93)*
Indigenous	0.141	0.090	0.137	0.091
	(5.04)***	(3.76)***	(4.81)***	(3.77)***
Community Match FE	x	x	x	x
R^2	0.03	0.04	0.02	0.02
N	12,667	12,667	12,667	12,667

* $p<0.1$; ** $p<0.05$; *** $p<0.01$

Table 15: IV Estimates of the Effect of HI on Health & Utilization

	Health Index (Follow-up)	No. HH Hospitalizations	Outpatient care (12 mth)	Hospitalization (12 mth)
Health Insurance	-0.007	-0.013	-0.058	0.001
	(0.98)	(0.59)	(.74)	(0.14)
Age	-0.001	-0.009	0.002	-0.005
	(2.44)*	(5.94)**	(2.06)*	(9.12)**
Age squared	0.000	0.000	-0.000	0.000
	(6.43)**	(4.77)**	(2.27)*	(7.75)**
Sex (male)	-0.013	-0.009	-0.173	-0.043
	(3.57)**	(0.78)	(24.74)**	(10.84)**
Educ: primary	-0.005	-0.000	0.028	-0.002
	(1.11)	(0.02)	(3.38)**	(0.46)
Educ: secondary	-0.009	-0.011	0.062	-0.002
	(1.91)	(0.75)	(6.71)**	(0.35)
Educ: college/vocational	-0.005	0.010	0.077	-0.003
	(0.92)	(0.61)	(7.67)**	(0.50)
Married	-0.006	0.033	0.066	0.014
	(2.06)*	(3.63)**	(11.30)**	(4.20)**
Dependents	-0.002	0.030	0.069	0.005
	(0.56)	(3.20)**	(11.68)**	(1.37)
Employed	-0.002	0.001	-0.010	-0.009
	(0.54)	(0.11)	(1.45)	(2.39)*
HH Size	0.000	0.013	-0.007	-0.003
	(0.26)	(5.69)**	(4.40)**	(3.39)**
Rural	0.006	-0.000	-0.042	0.017
	(1.16)	(0.03)	(4.35)**	(3.06)**
Indigenous	-0.014	-0.047	-0.047	-0.004
	(2.42)*	(2.68)**	(4.24)**	(0.60)
Oportunidades	-0.005	-0.037	0.060	-0.009
	(1.50)	(3.90)**	(10.08)**	(2.77)**
R^2	0.03	0.01	0.04	0.02

* $p<0.1$; ** $p<0.05$; *** $p<0.01$

Table 16: Correlation between Health Changes and Preventive Behaviors

	Flu Vaccine, Baseline	Smoker, Baseline	Eye Exam, Ever	Eye Exam, Baseline
Change in Index	0.030	-0.014	0.080	0.040
	(1.91)*	(1.17)	(4.30)***	(3.37)***
HH Size	-0.011	0.002	-0.016	-0.005
	(8.41)***	(2.39)**	(10.33)***	(4.96)***
Rural	-0.003	0.016	-0.227	-0.074
	(0.09)	(0.65)	(5.92)***	(3.03)***
Indigenous	0.041	-0.005	-0.069	-0.035
	(3.24)***	(0.48)	(4.61)***	(3.61)***
Oportunidades	0.070	-0.009	-0.045	-0.003
	(12.18)***	(2.00)**	(6.56)***	(0.64)
Community Match FE	x	x	x	x
R^2	0.04	0.02	0.06	0.02
N	24,956	24,956	24,956	24,956

* $p<0.1$; ** $p<0.05$; *** $p<0.01$

Figure 1


Sonora
San Luis Potosí
Estado de México
Jalisco
Guerrero
Oaxaca
Morelos

Figure 2

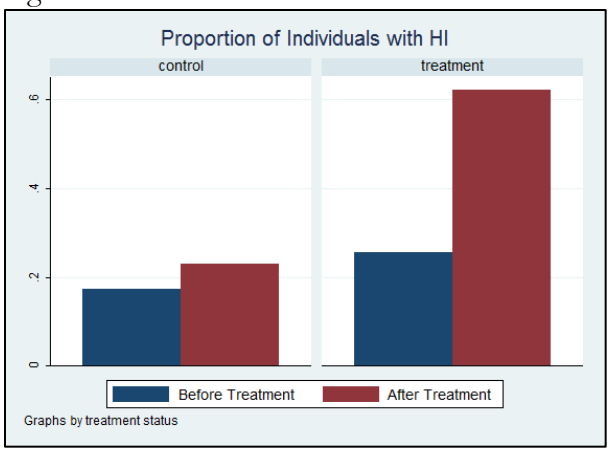


Figure 3

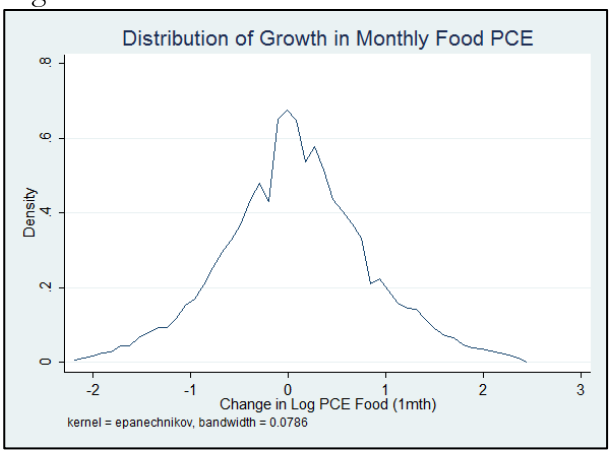


Figure 4

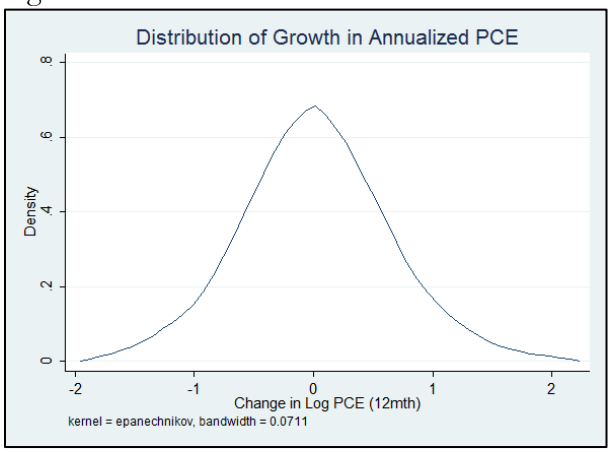
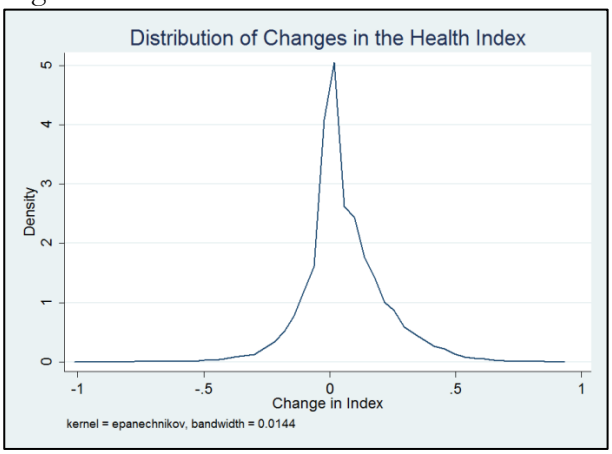


Figure 5



The Effectiveness of Environmental Alerts: Evidence from Santiago, Chile

Jamie Mullins* Prashant Bharadwaj†

January 20, 2013

Abstract: Through the 1990's, the government of Chile implemented various measures to reduce air pollution in its capital city, Santiago. This paper focuses on examining short and long term effectiveness of a particular policy of issuing environmental "alerts" when ambient air pollution reaches dangerously high levels. We find that these policies are correlated with lower air pollution and air pollution related deaths in the Santiago Metropolitan Area over the long run. Using propensity score matching and difference in differences, we show that Chile's Environmental Episodes approach to addressing severe air pollution events in the short run is quite effective, with the announcement of an Environmental Episode leading to reductions in PM_{10} concentrations on the order of 30%. While we lack power to relate the short term effects of Episode announcements to death rates, the point estimates of the effects are consistent with the idea that episode announcements mitigate the incidence of negative health outcomes associated with exposure to particulate matter.

Keywords: Pollution alerts, particulate matter, mortality

*Department of Economics, UC San Diego
†Department of Economics, UC San Diego

1 Introduction

Over the past 50 years, an increasing number of national, regional, and city governments around the world have taken deliberate actions to address local air pollution. Unlike global pollution issues such as the deterioration of the Ozone layer or global temperature rises, local air pollution problems arise due to location specific economic development, topography and weather patterns. Similarly, government responses to local air pollution have varied greatly in their methods and levels of success. This paper will focus on the policies implemented by the Chilean government to combat the high levels of ambient air pollution that have plagued Santiago since at least the 1960's, when the pollution monitors were put in place.

Santiago is particularly susceptible to elevated levels of a number of air pollutants due to the weather patterns and topographical characteristics of its location. This paper will follow the Chilean government's focus on PM_{10}[1] as a criteria pollutant, and thus our discussion of pollution levels will be anchored by measured levels of PM_{10}. Prior to the start of significant government interventions, PM_{10} levels within the city of greater than $300\mu g/m^3$ were not uncommon and occasionally levels of greater than $500\mu g/m^3$ were measured. Considering that World Health Organization guidelines for PM_{10} are currently set at a 24-hr mean value of $50\mu g/m^3$ it is unsurprising that Santiago was widely known for its poor air quality (WHO, 2011).

Exposure to air pollution has been shown to impose societal costs through a number of different channels[2], however the most salient cost of air pollution is the negative impacts it has on the health of exposed populations. A large amount of research (both economic and epidemiological) has linked exposure to air pollution to negative health outcomes (Dockery et al., 1993) with a large fraction of this literature focusing specifically on the negative effects of particulate matter exposure on health in both the short (Dominici et al., 2003, Katsouyanni et al., 2001) and the long term (Pope et al., 1995, 2002).While the most commonly cited health effects of exposure to air pollution are related to respiratory ailments, research has shown that elevated levels of ambient particulate matter also significantly impact outcomes of cardiovascular disease among exposed populations (Pope et al., 2004). Additionally, there have been several studies which consider the timing of the negative effects of exposure to air pollution. Zanobetti et al. (2003) find that the increased mortality rates that correspond to elevated levels of particulate matter are not due solely to a net harvesting effect,[3] while Goodman et al. (2004) concluded that increases in

[1] PM_{10} is a measure of the concentration of airborne particulates smaller than or equal to 10 micrometers in size in a given volume of air. It is commonly reported in micro-grams of particles per cubic meter of air.

[2] For example: degradation of asset valuations (Chay and Greenstone, 2005), diminished quality of life (Luechinger, 2009), reduced experiential values (Carson et al., 1992), property and ecosystem damage (Likens et al., 1996), and reductions in economic output (Ostro, 1983, Zivin and Neidell, 2011).

[3] The hypothesis of the Harvesting Effect holds that while exposure to high concentrations

mortality due to elevated levels of particulate matter may persist for days or even weeks after a exposure.[4]

Such results, along with growing public discontent, motivated a string of actions by the Chilean government through the early 1990s to address the worsening air pollution in Santiago. These actions culminated in 1997 with the publication of the Plan de Prevención y Descontaminación Atmosférica (hereafter "PPDA"),[5] which laid out a region-wide governmental approach explicitly intended to reduce air pollution in the Santiago Metro Region and mitigate the negative health effects of air pollution exposure among the population. While the PPDA included a number of provisions, chief among them was the governmental issuance of tiered environmental alerts on days for which levels of air pollution exceeded (or were expected to exceed) certain threshold levels. Announcement of the higher levels of such "Environmental Episodes" (as they are called) were accompanied by mandatory restrictions on driving as well as the shut down of certain major stationary emitters. The analysis in this paper focuses on the effectiveness of the entire set of policies that accompanies such Episode announcement, rather than a specific component of the various restrictions.

One of the main challenges in an empirical examination of such policies is isolating the impact of the policy from other factors that might also be driving pollution levels. Moreover, when examining health impacts, avoidance behavior becomes an important confounder (Neidell, 2009, Moretti and Neidell, 2011). As a first pass, we examine air quality and mortality data from before and after 1997 in Santiago and find substantial changes in air quality and reductions in air quality related mortality. Since this does not imply a causal relationship, we turn to examining the short run effectiveness of these policies within a framework that allows us to more accurately compute counterfactuals. Along with avoidance behavior, a more basic issue in this context is mean reversion. Since alerts are announced when pollution levels typically reach their maximum, it is hard to pin down whether the subsequent drop is due to the policy or simply the recession from a natural maximum which would have occurred regardless of policy actions. While it is difficult to pin down the causal impacts of such policies in the long run, we can do better in the short run by using the fact that episodes are only announced on certain days when pollution levels reach a certain level. By comparing Episode days to days with similar pollution levels but with no Episode announcements we are able to ascertain the effectiveness

of air pollution lead to increases in the number of deaths observed, these deaths are likely occurring among an enfeebled population that was likely to die soon in any case. Further, the Harvesting Effect requires that this pool of enfeebled and infirm is not expanded by high air pollution events to the same extent to which it contracts through hastened deaths.

[4] For the purposes of this paper, it is also worth making note of a number of studies, which examine the health costs of air pollution in Santiago itself. Several of these papers demonstrated a positive relationship between elevated levels of ambient particulate matter and mortality rates (Ostro et al., 1996, Salinas et al., 1995, Cifuentes et al., 2000, Sanhueza H et al., 1999), and a number of others found correlations between high levels of particulate matter and the incidence of infant and childhood respiratory ailments and hospital visits (Pino et al., 2004, Ilabaca et al., 1999, Ostro et al., 1999).

[5] PPDA translates as the Plan to Prevent and Reduce Air Pollution.

of Episode announcements in reducing pollution levels and air quality related mortality. Moreover, we exploit the fact that the Episode announcements were not fully implemented before 1997, giving our design a period of time when such policies were not in place. Our empirical methodology consists of a combination of propensity score matching and difference in differences, which does not resolve the issue of avoidance behavior, but instead identifies the full impact of the policy on mortality.

We show that overall, Chilean policies have led to large improvements in air quality and related health outcomes in the Santiago metropolitan area since the early 1990s. In the 5 years after the 1997 implementation of the PPDA, both the average daily PM_{10} levels and the rate of respiratory related deaths among the elderly declined by approximately 25%. In addition to these long term improvements, we find that Chile has been able to effectively address high-PM_{10} levels on a short term basis due to the use of Environmental Episode announcements. Days after an Episode announcement experience significantly lower levels of pollution compared to similar days with no announcement. For example, day 2 after an Episode announcement has 21% percent lower PM_{10} levels. Hence, in both the short and long term, the policies introduced by the PPDA appear to have had a significant impact in improving air quality in Santiago. However, we lack power to make causal statements about mortality effects in the short run, even though the coefficients indicate that better air quality results in fewer mortality numbers.

Our paper adds to the literature measuring the effects of localized governmental policies aimed at curbing air pollution. Most of these studies and policies focus on driving restrictions and have found mixed results. Using data from São Paulo, Bogotá, Beijing, and Tianjin, Lin et al. (2011) find that driving restrictions may reduce the incidence of extremely high concentrations of air pollutants, but that such restrictions do not, on average, significantly improve overall urban air quality. Conversely, Viard and Fu (2011) find that every-other-day restrictions in Beijing (implemented prior to and during the 2008 Olympic Games) reduced total air pollution levels by 19%, while subsequent one-day-per-week restrictions (which have continued since 2008) have lead to an 8% overall reduction in air pollution. Additionally, Viard and Fu find that driving restrictions seem to reduce hours worked for the self-employed, hinting at the possible unintended negative effects of such pollution reduction programs(Viard and Fu, 2011). Davis (2008) concludes that one-weekday-per-week driving restriction in Mexico City did not significantly improve air quality in that city, and postulates that the restriction likely drove the purchase of additional (often older, dirtier) vehicles by households to get around the restrictions. In addition to examining the impacts of such policies on pollution levels in Santiago, we extend our analysis to examine the health impacts of such policies. Quantifying this natural externality of pollution abatement policies is a novel addition to the literature.

As this paper addresses the short-term effectiveness of Episode announcements in Santiago, it is important to note the paper by Troncoso et al. (2012) which finds that the announcement of Environmental Episodes leads to significant reductions of particulate matter, CO, NO_x, and O_3, while having no effect

on SO_2. In addition to an examination of the health impacts of the Santiago policies, this paper adds to the Troncoso et al. (2012) analysis in a number of other important ways. First, a broader time period, including the initial introduction of the PPDA is examined, allowing us to take advantage of greater overall variation in pollution. Second, we explicitly account for mean reversion of ambient pollutant levels and the fact that Episode announcements are likely to correspond with natural peaks in pollution levels. Third, we address the mutli-day effects of Episode announcement, which are not discussed by Troncoso et al. (2012) but prove to be a significant portion of the overall benefit of such announcements. Finally, the approach of this paper allows for the examination of the treatment effects of Episode announcement, while Troncoso et al. (2012) are restricted to characterizing pollution levels contemporaneous with Episode announcements.

2 Air pollution alerts in Santiago

Santiago sits in a basin at the western base of the Andes mountain range,[6] which leads to the frequent occurrence of temperature inversion layers over the city. Such inversion layers reduce vertical atmospheric mixing, thereby trapping pollutant emissions near the ground. Thermal inversions are common in summer as well as in winter; however, in the winter months (April-August), the inversion layers tend to be much closer to the ground, leading to an increased intensity and prevalence of high-PM_{10} concentrations(Gramsch et al., 2006, Rutllant and Garreaud, 1995). Since the 1960's Santiago has periodically suffered from periods of air pollution far in excess of levels considered healthy for the region's ever growing population (now estimated at 7 million people).

Beginning in the late 1980s, the government of Chile implemented a string of policies intended to address air pollution in the greater Santiago area. This included the establishment of an automated network of pollution monitors in 1988, and, in 1990, the mandated inspection and ranking of stationary emitters based on the concentration of pollutants emitted (Chilean Ministry of the Environment, 2007). Also in 1990, the Chilean Government set up a system of Environmental Episodes through which announcements of a Pre-Emergency Episode were made for days expected to have PM_{10} concentrations in excess of $240\mu g/m^3$, and announcements of Emergency Episodes were made on days with expected PM_{10} levels over 330 $\mu g/m^3$(Supreme Decree 32,1990).

Along with the identification of days with potentially severe levels of air pollution, Episode announcements also carried with them mandated restrictions. Pre-Emergency Episodes required that the dirtiest 20% of stationary emitters[7] temporarily shut down, while Emergency Episodes required the dirtiest 50% to cease operations(Schreifels, 2011).[8] However, even though the policy of

[6] Santiago is also flanked to its West by the significant Coast Range, to the North by the Chacabuco mountains, and to the South by the Cantillana range Prendez et al. (2011).

[7] As ranked based on the required annual inspections.

[8] See Table A.1 for a full list of restrictions by Episode level.

announcing episodes was established in the early 1990s, its implementation was far from perfect. As we elaborate below, the implementation of environmental episodes improved substantially after the passage of the PPDA. Pre-PPDA, given the thresholds of when an announcement should have been made, we can compare actual data on PM_{10} and compare it to data on when announcements were made. Out of potentially 151 days when $PM_{10}$was in excess of $240\mu g/m^3$ less than 40% of days were announced as Episodes. After the passage of the PPDA, nearly all days when a pollution should have been called were days when an announcement was made. We conclude that the policy of environmental alerts only really came into force after 1997.

Other significant pollution abatement measures in the early 1990s included a cap-and-trade system for stationary polluters (Schreifels, 2011), and a requirement that all new cars be fitted with a catalytic converter(Bauner and Laestadius, 2003), both implemented in 1992. Then in 1994, it was mandated that CONAMA (the National Commission on the Environment) develop a comprehensive approach to cleanup and control Santiago's air pollution. The result of these efforts was the PPDA, the implementation of which began in 1997 (though not officially signed by the president until 1998) (Chilean National Environmental Commission, 1998). It should also be noted that throughout the 1990s, stationary emitters had to conform to a series of increasingly stringent regulations regarding the total volume of potential emissions (Schreifels, 2011). However, none of these regulations was directly part of the PPDA enforcement.

The PPDA included a wide range of policy measures intended to reduce concentrations of air pollution in the Santiago Metropolitan Area and to mitigate the negative health impacts of air pollution in the city (Chilean Ministry of the Environment, 2007). Given the numerous policies implemented in Santiago during the early and mid-1990s, and the "bundle" of policies included in the PPDA, it is likely impossible to identify the impact of any one policy or action in the long term. However, a major piece of the PPDA was an update and better implementation of the Environmental Episode system originally implemented in 1990. While we show some suggestive evidence of the long-term success of Chile's air quality improvement campaign as a result of all the policies under the PPDA, the econometric analysis will be focused much more narrowly on the effectiveness of the revised Environmental Episode program in meeting the goals of the PPDA in the short term.

Under the provisions of the PPDA, just as under the original implementation of the Environmental Episodes program, each episode level was accompanied by a prescribed set of actions and restrictions which were implemented automatically upon announcement. As the announcement and protocol implementation happen coincidentally, it will not be possible to attribute effects separately, thus we will empirically examine the combined impacts of the announcement itself and the prescribed government actions which the announcement invokes.[9] For a

[9] The PPDA also introduced a new level of Environmental Episode implemented when PM_{10} concentrations were expected to be $> 195\mu g/m^3$. The new level is called an "Alert" Episode. This level is generally omitted from the analysis of this paper because it came with only minimal additional driving restrictions and no implications for stationary emitters. The

complete list of protocols by Episode level, please see Table A.1 in the appendix of this paper. A key aspect for our empirical design is that there were no changes to the thresholds leading to alert announcements under the PPDA.

Two of the most significant changes to the Environmental Episodes program that were made under the PPDA after 1997 were actually updates to strategies that had been in place previously. Although seasonal driving restrictions had been in place since the late 1980s, restricting 20% of cars from the road on each weekday from April - August, the PPDA added additional driving restrictions which came into force upon the announcement of a Pre-Emergency or Emergency level episode. Specifically, when a Pre-Emergency was announced on a weekday, 60% of cars without catalytic converters and 20% of cars with catalytic converters were banned from use. The restrictions increase to 80% and 40% for an Emergency Episode. Table 1 outlines the share of vehicles that were restricted at any given time under the PPDA.

The driving restrictions under the PPDA remained unchanged until 2007. At that time the seasonal restrictions were raised to 40% and the weekend restriction was dropped for non-catalytic converted cars. Also, the restrictions for automobiles with catalytic converters were increased from 20% and 40%, to 40% and 60% in the events of Pre-Emergency and Emergency Episodes respectively.

Table 1. Percent of Vehicle Fleet Restricted under PPDA, by Season and Episode Level

	Vehicles with Catalytic Converter		Vehicles W/O Catalytic Converter	
Episode Level	Weekdays	Weekends	Weekdays	Weekends
Seasonal (Apr.-Aug. Restrictions)	0%		20%*	0%
Alert	0%		40%	20%**
Pre-Emergency	20%***		60%	40%
Emergency	40%***		80%	60%

Notes: Data for the table is taken from the PPDA (Supreme Decree No. 32, 1998; Supreme Decree No. 46, 2007).

Once set under the PPDA, vehicle restrictions were not changed until 2007. At that time, the three starred figures were updated as noted below:

*- changed to 40% in 2007

**- changed to 0% in 2007

*** - increased by 20% in 2007

In addition to the vehicle restrictions, the PPDA updated the stationary emitter shutdown requirements connected to Environmental Episode announcements. Under the PPDA, stationary emitters were still inspected each year, and ranked on their emissions. However, rather than a fixed share of stationary

protocols that come into force upon the announcement of an Alert Episode are summarized in Table A.1.

emitters being shut down as was done previously, under the PPDA emitters were forced to shut down only if their emissions exceeded certain concentration thresholds. On days for which Pre-Emergency episodes were announced, facilities that had not been certified to have emissions levels less than $32mg/m^3N$ had to shut down their emitting operations, and when an Emergency was announced, facilities that had not be certified to have emissions levels below $28mg/m^3N$ were shut down.

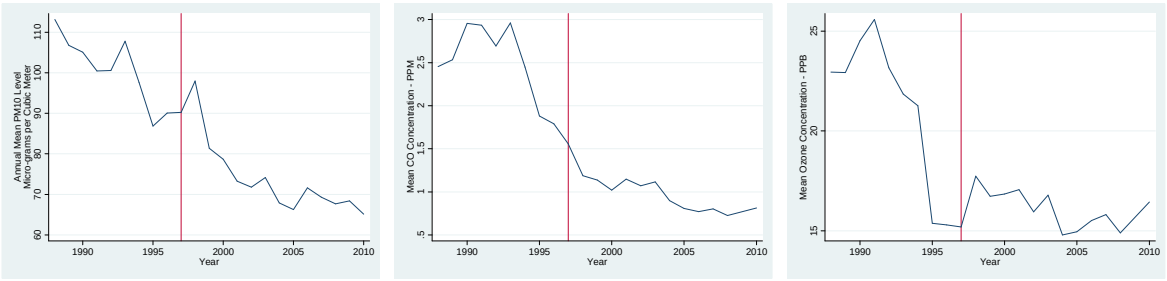
2.1 From policy to impact: suggestive evidence

Since the late 1980s, air quality has been improving generally in the Santiago metro area. Regular data is not available far enough back to observe the turning point from increasingly worsening air pollution to the improvements that we see more recently. Although the goals of the PPDA included reduction of a number of air pollutants, the plan was explicitly implemented to address PM_{10}, CO, and $Ozone$. Panel A of Figure 1 shows the yearly mean PM_{10} concentrations across all active monitoring stations for the period from 1988-2010. We see that PM_{10} levels are generally falling throughout the period, and Panel B of Figure 1 shows that Carbon Monoxide follows a similar downward pattern.

Conversely, we see in Panel C of Figure 1 that Ozone levels within Santiago fell sharply during the early 1990s, but have essentially leveled off since 1995. It seems clear that the PPDA's introduction in 1997 (marked by the red line on the graphs) was not associated with significant reductions in Ozone levels. This may be due to the central role that environmental episodes play under the PPDA, and the way in which episodes are announced. Ozone levels are highest during the summer months in Santiago, when PM_{10} levels tend to be lower. Since episodes are announced based on PM_{10} levels, it is likely that events of high-Ozone concentrations within the city go unaddressed under the PPDA. What we do see in the graph seems to suggest that earlier measures taken by the Chilean government seem to have lead to significantly reduced levels of Ozone in the city. Notably, as mentioned previously, in 1992 catalytic converters were required for all new vehicles and a ratcheting set of emissions restrictions was put onto stationary emitters. These restrictions on stationary emitters increased their stringency each year through 1997 (Schreifels, 2011). This time frame coincides very closely with the fall in Ozone levels depicted in Panel C of Figure 1, but there does not seem to be any additional impact of the PPDA on average Ozone levels in Santiago.

Reductions in PM_{10} appear to have been due to citywide improvements in air quality throughout the 1990s as stated in the previous section. With the implementation of the PPDA , a new network of pollution monitors was also established. The 6 monitors in this so-called MACAM2 network replaced the 5 monitors in the MACAM network. In contrast to the original MACAM network, the MACAM2 monitors are spread deliberately throughout the city, with placements intended to capture traditional hotspots and provide observations on

Figure 1. PM_{10} **Annual Mean Concentrations: 1988-2010**



(a) PM_{10} **(b)** Carbon Monoxide **(c)** Ozone

Notes: Red line marks the start of 1997, when the PPDA implementation began. All plots show yearly mean values across all days and all monitors for which data was available. Pollution data source is Chile's Ministry of the Environment. Pre-1997 data comes from the 5 monitors in the MACAM monitoring system. Starting in 1997, data is from from 6 monitors in the MACAM2 monitoring network initiated under the PPDA. The number of Monitors in the MACAM2 network grew to 8 by 2008. Please see Figure A.1 for maps showing the locations of the monitors in each network.

pollution levels actually experienced by the population(Gramsch et al., 2006). The data show falling average PM_{10} levels at all monitors throughout the city beginning in the late 1980s.[10] This trend appears to continue across the city through 2005 when PM_{10} concentrations level out generally, as can be seen in Panel A of Figure 1.

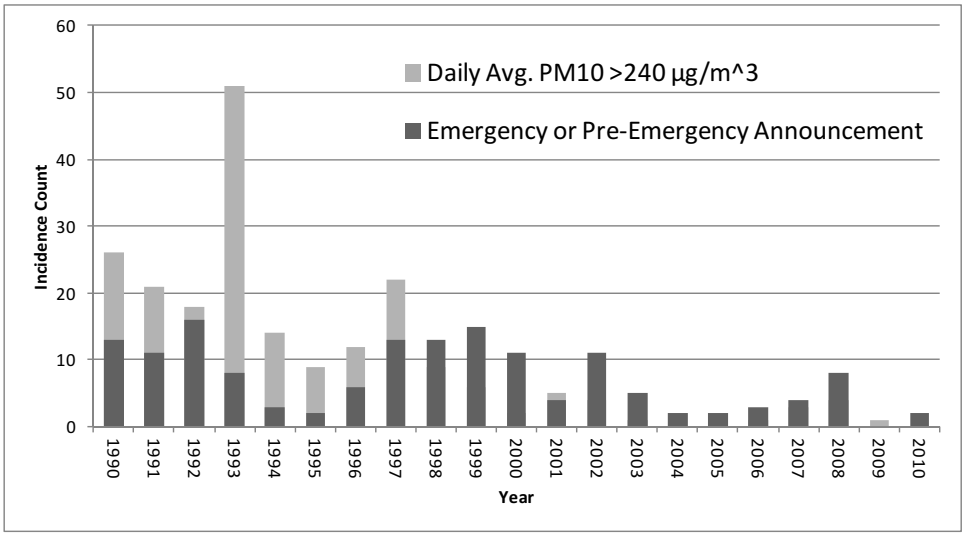
Another way to assess possible improvement in Santiago's air quality is to examine the number of environmental episodes announced each year. Figure 2 shows the number of environmental episodes (ignoring the "Alert" days, and just focusing on Pre-Emergency and Emergency Episodes) announced in each year against the backdrop of number of high PM_{10} days each year. What is significant about 2 is that before 1997, only some of the high pollution days were identified as days when an Episode announcement was made. After 1997, it is clear that almost all high pollution days were identified as Episode days. The results in 2 do not seem immediately consistent with falling average-PM_{10} levels over the period of the graph. In fact, further examination of the data suggests that the two dip/spike events in Episode announcements were likely the result of changes in government behavior rather than changes in pollution levels. The first dip in Episode announcements reached its nadir in 1995, and significantly bounced back coincident with the introduction of the PPDA and thereafter.[11] The second bounce in episode announcements began in 2006, which was the year in which the results of a major audit of the PPDA plan were released

[10] See Figure A.2 for PM_{10} levels by station.

[11] Note, the apparent increased vigor of Episode announcement in 1997 and following is even more apparent if the Alert level Episodes are included.

(Lents et al., 2006).[12] The evidence indicates that both upticks in Episode announcement occur following events which brought attention to Santiago's air pollution, suggesting that these upticks may have been related to additional government attention rather than an underlying change in pollution levels. To drive this point home, note the similarities between 1996 v. 1998 and 1995 v. 1999 in Figure 2. Each pair had comparable numbers of high-PM_{10} days, yet we see in Figure 1 that the years after the implementation of the PPDA (namely: 1998 & 1999) have dramatically more Pre-Emergency and Emergency level episode announcements. Thus, we see years with similar air pollution profiles meeting with quite different government actions, suggesting that the vigilance with which episode announcements were made was different before versus after the implementation of the PPDA. A similar, although somewhat less compelling, case can be made for the differences in episode announcements between the period before the early-2006 audit release and the period after the release by looking at 2003 v. 2007 and 2004 v. 2006. Thus we see that although government vigilance fluctuated over the period of interest, the overall number of severe PM_{10} events has been falling fairly consistently at least since the PPDA came into force.

Figure 2. Episodes and High-PM_{10} Events by Year



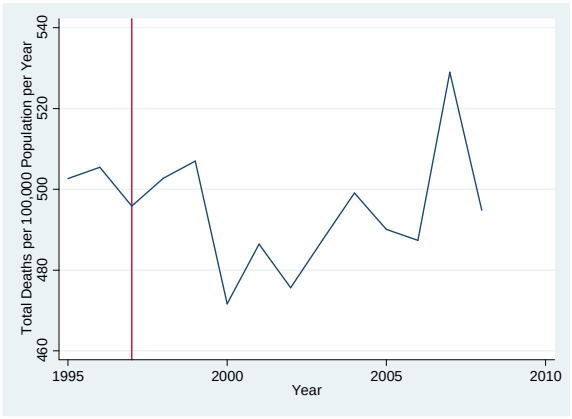
Notes: Data on historical episode announcements are available from the Metropolitan Region Ministry of Health. Data for PM_{10} levels are available from the National Ministry of the Environment.

The positive air quality trends in the data suggest a general effectiveness of the Chilean policies. However, we have so far looked only at levels of air

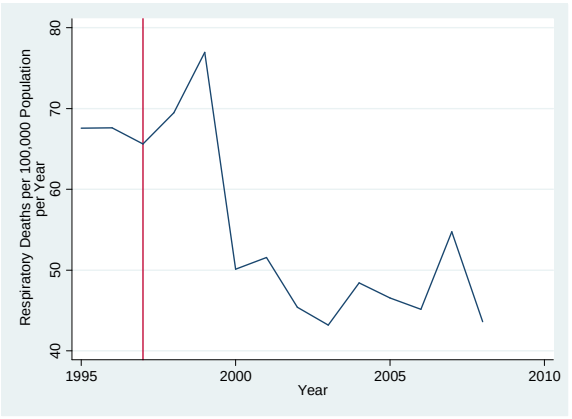
[12]The results of the audit lead to the 2008 update to the PPDA, which was officially rubber stamped in 2009.

Figure 3. Deaths by Year

(a) Total Deaths per 100,000



(b) Respiratory Deaths per 100,000



Notes: Death counts are aggregate statistics obtained by summing reported deaths of a given type for each day. The data identifies a primary and potentially a secondary cause of death for each observation in data using the ICD-10 classification system. Deaths were identified as "Respiratory" only if the primary cause of death was attributed to ailments of the respiratory system. Data on deaths was obtain from Chile's Ministry of Health.

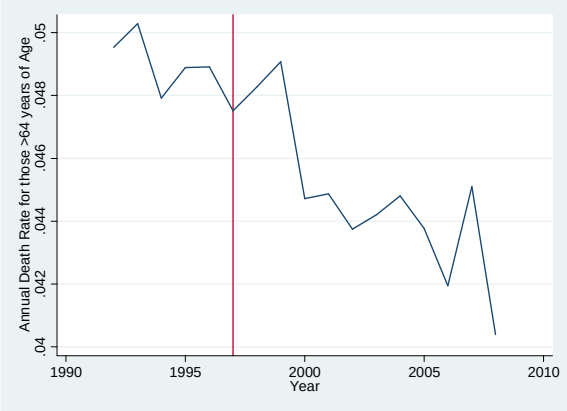
pollution, but with the advent of the PPDA, the Chilean government expanded the explicit goals of its air pollution policies to include the mitigation of negative health outcomes related to air pollution. In order to examine long run health impacts of Chilean air pollution policies we will look at death rates within the Metropolitan Region over the period from 1992-2008.

Panel A of Figure 3 plots total annual deaths (all causes and ages) per 100,000 people by year. This very simple representation does not suggest that the death rate in greater Santiago has obviously diminished since Chile's string of air pollution policies were enacted in the early to mid-1990s. However, if we instead examine deaths attributed primarily to respiratory ailments, we see a very different story. Panel B of Figure 3 shows a sharp decline in the annual number of respiratory deaths per 100,000 people in Santiago that tracks closely to the falling mean levels of PM_{10} within the city shown in Panel a of 1. These results fit nicely with the associations between respiratory ailments and exposure to particulate matter reported in the literature(Dockery and Pope, 1994).

Looking now at death rates within age delineated sub-populations in Figure 4, we see that health outcomes for both the elderly and the very young are improving over the period during which Santiago air quality was dramatically improving. Though such co-moving trends are not more than suggestive evidence of some impact of air pollution policy (presumably improvements to neonatal or elderly care during this period could have produced the same results), the literature has found clear connections between air pollution and the health of

Figure 4. Deaths by Age

(a) Death Rate: Age > 64 years



(b) Deaths Rate: Age < 5 years

Notes: Data on deaths by age was obtained from Chile's Ministry of Health. Age range death rates are calculated by dividing the total number of deaths in an age range by the estimated total population of that age range in the Metropolitan Region. Age range population figures are linear interpolations from 5 year estimates published by Chile's National Institute of Statistics based on the 2002 National Census.

these sub-populations (Saldiva et al., 1995, Ilabaca et al., 1999, Ostro et al., 1999).

While no causal claims can be made from this qualitative line of investigation, the evidence clearly demonstrates that significant reductions in PM_{10} and CO levels accompanied the roll out of Chile's air pollution control policies. Additionally, there are clearly positive changes in health outcomes of the young and the old in Santiago coincident with the falls in pollution levels. Taken all together, these observations suggest a certain level of long run success for Chile's efforts to reduce air pollution in Santiago and mitigate the damaging health impacts on its population. However, for analysis within a causal framework, we turn to analyzing the short run implications of episode announcements.

3 Empirical Approach and Data

In order to examine the short term effects of high-level (Pre-Emergency or Emergency) Episode announcements we will have to develop an approach that allows us to control for mean reversion. Because Episodes are typically announced at times when air pollution is above mean concentrations, on average we would expect air pollution levels to fall on subsequent days whether or not any actions were taken. Failing to take account of this fact will lead to upward bias in our estimated effects. Additionally, other factors, including weather shifts, may lead to differential changes in pollution levels that are unrelated to Episode

announcements. In order to account for mean reversion (and perhaps other effects that might change outcomes but are unrelated to Episode announcements) we will utilize a difference-in-differences approach, comparing changes in outcomes from before to after an Episode to changes in outcomes from before to after another day when no Episode was announced. The following section will focus on identifying the proper group of proxy-counterfactual days for Post-PPDA Episode announcements, and the subsequent section will outline the Difference-in-Differences approach we use for this analysis. Figure 3 explains in intuitive terms, why such methods are needed in this case.

3.1 Identifying Appropriate Counterfactuals

Since we are particularly interested in the short-term impacts of Episode announcements under the PPDA protocols (i.e.- 1997 and after), an obvious comparison group is the days on which Episodes were announced before the PPDA. While this approach offers a simple and intuitively appealing way to identify the the effects of the PPDA updates to the Environmental Episodes program, we showed previously that Episode announcement criteria were not being consistently applied during the several years preceding the implementation of the PPDA. This point is made particularly clear by examining Figure 2. Thus we dismiss government announced Episodes in the Pre-PPDA period as a useful comparison group for Post-PPDA Episodes since the announcements lacked consistency.

Given that the government identified Episodes cannot serve as a valid counterfactual for the Episodes of interest, we must identify another set of days for which we would expect comparisons to yield informative results. Intuitively, these are days during the Pre-PPDA period when an Episode should have been announced, but were not announced. Propensity score matching provides a quantitative approach to matching Post-PPDA-Episode days to similar days on which an Episode of interest was not announced. As we have done throughout this paper, we focus only on the Pre-Emergency and Emergency Episode levels.[13] This approach was chosen because it provides greater separation between Episode announcements, allowing for better identification of treatment effects, and because the restrictions and protocols which come into effect under the Pre-Emergency and Emergency level Episodes are much more severe, making Episode impacts easier to detect. Furthermore, our analysis is conducted only on Episode announcements which were neither preceded nor followed by another Episode announcement within 5 days.[14] This leaves us with a sample of 36 Episodes which will be the focus of our analysis.[15]

Each Episode in the Post-PPDA period is matched, based on a Probit

[13] Hereafter, all usages of the word "Episode" will refer only to Pre-Emergency and Emergency level Episodes.

[14] That is, there are at least 4 non-Episode days between each Episode we define as "stand alone" and use in our analysis.

[15] Note that the presented results are robust to the use of a variety of other sub-samples of Post-PPDA Episodes. See section 4.1 for further discussion.

generated propensity score, to a number of days in the Pre-PPDA.[16] The propensity score can be thought of as an estimated probability that a given day would have had an Episode announcement. The value of the score is generated by running a Probit model on predetermined characteristics of days in the Post-PPDA period on which Episodes were announced, and using the estimated coefficients to predict the probability that a given day in the Pre-PPDA would have had an Episode announcement. The Probit model is first estimated on the Post-PPDA data using the following specification, and the results of the Probit model estimation are presented in Table A.3.

$$y_t = \alpha + \beta_1 * PM10_{t-1} + \beta_2 * PM10_{t-2} + \hat{\delta} * \mathbf{DOW_t} + \hat{\theta} * \mathbf{month_t} + \hat{\gamma}_1 * \mathbf{X_{t-1}}$$

$$+\hat{\gamma}_2 * \mathbf{X_{t-2}} + \hat{\gamma}_3 * \mathbf{X_{t-3}} + \hat{\gamma}_4 * \mathbf{X_{t-4}} + \varepsilon_t$$

where:

- y_t : indicator variable =1 if an Episode was announced on day t
- $PM10_t$: Mean, citywide PM_{10} concentration on day t
- $\mathbf{DOW}_t$: vector of dummy variables for each day of the week
- $\mathbf{month_t}$: vector of dummy variables for each month of the year
- X_t : vector of observed weather variables on day t including mean temperature, average windspeed, dew point, the square of mean temperature, and the cube of mean temperature
- ε_t : error term

A few things are worth noting about the chosen model specification. Only two lags odata on PM_{10} and compare it to data on when announcements were n pollution levels were used due to findings from Saide et al. (2011) showing that Santiago PM_{10} levels on day t are only a function of emissions from days t, $t-1$, and $t-2$ (Saide et al., 2011). Adding more lags does not alter our results meaningfully, although a slight loss in precision occurs as a result of having a smaller support. Weather variables for day t are excluded because their levels are not determined at the time of the Episode announcements. Day of the week was included as it likely captures some emissions information that may improve match quality.[17] Month dummies were included to capture both seasonal variation in weather patterns which likely contribute to fluctuations in

[16] We allow days in the Pre-PPDA period on which Episodes were announced to be matched to ensure the best possible matches. If these Pre-PPDA Episodes did have positive effects on the outcomes of interest, the inclusion of these days will bias our estimates downward.

[17] Additionally, although it is not technically a factor that should be considered by authorities, the day of the week may impact the government decision of whether to announce and Episode or not.

air pollution, and seasonal variation in Episode announcements connected to attitudes within government.

Once the above Probit model is estimated, each Post-PPDA Episode and Pre-PPDA day is plugged into the model, and the resulting $\hat{y}_t$is the propensity score for day t. Each Post-PPDA episode is then matched to, possibly multiple, Pre-PPDA days based on the radius method using a caliper of width=0.05.[18] Matching each Episode to multiple Pre-PPDA days reduces the variance of our estimates, while limiting the matching radius reduces the possibility of using poor matches in our analysis.

3.2 Difference-in-Differences

Now that we have identified a set of matched days in the Pre-PPDA period, we have an appropriate "control" group against which to compare the outcomes of our "treatment" group of days with a Post-PPDA Episode announcement. Since the goal of this exercise is to identify the effects of an Episode on several different variables, the outcome variables we will compare between the treatment and control groups will be differences over time, across the date of the Episode (or Episode match). We are specifically interested in the effects that Episodes (and the accompanying mandated government actions) have actually had when they have been announced.

Our first difference will be changes in pollution levels or death rates (our outcome variables of interest) from before to after an Episode (or matched day). If everything were held constant but the Episode announcement, this change in the outcome variables would be our treatment effect. However, we know that many other factors that impact air pollution and/or death rates are likely changing contemporaneously to each Episode announcement (weather being the most obvious of these factors), therefore we control for these factors by subtracting the change in our outcome variable from before to after our matched day from the change in our outcome variable from before to after an Episode. The resulting difference in differences identifies the treatment effect on the treated as the amount by which changes in the outcome variable are larger with treatment than without. No additional controls are needed at this stage of the analysis, because the propensity score matching effectively controls for all the inputs which are matched over.

3.3 Data

The empirical analysis of short run effects in this paper relies on a data set created by merging a number of Chilean government-provided data sources. At the very core of this examination are observed PM_{10} concentrations from the MACAM (pre-1997) and MACAM2 (post-1997) monitor networks. These networks also collected observational data on CO and $Ozone$ levels, as well as several weather variables for the period from 1988-2008. All of this data is

[18]This method involves each Post-PPDA episode being matched to all Pre-PPDA days with propensity scores within 0.05 of the propensity score of the Post-PPDA episode.

available publicly and was obtained for this project from the website of Chile's Ministry of the Environment. Additionally, we added data on number of deaths per day by age and cause of death. Mortality data is available publicly at the anonymous person level back to 1994 from Chile's Ministry of health and was obtained from 1992 and 1993 from the Chilean government, and we aggregated deaths up to the level of a daily death count, and include projected population numbers for the Ministry of Statistics for the calculation of mortality rates. Finally, to our daily values of pollution, weather, and deaths we add dates and levels for government announced Environmental Episodes. This data is available from the Metropolitan Region's Ministry of Health and covers the time period from 1990 (when the Environmental Episode program was first implemented) to 2008. Table A.4 is a table of mean values for the variables used in the analysis.

4 Results

Panels A and B of Figure 5 show the average movement of air pollution and mortality through time across the threshold of an Episode announcement. In each of the graphs, the red vertical line represents the announcement of an Episode, while the horizontal axis groups days by the amount of time before or after an Episode that they occur. As we would expect if the Episode program were successful in the short-term, we see both PM_{10} and deaths climbing steadily before Episode announcement, and responding quickly and beneficially upon Episode implementation As discussed earlier, in these graphs, mean reversion is a critical concern.

The first step to estimating the short-term treatment effect of Episode announcements under the PPDA is to appropriately match such events to comparable days in the Pre-PPDA period. Overall, we are able to provide at least one valid match for all 35 of the 36 Post-PPDA Episode announcements that meet our selection criteria.[19] Once this is done, in the next step, we calculate the changes across the Episode or matched day for the treatment and control groups separately.[20] The difference between these changes is our difference in difference estimate. Table 2 reports the mean difference between the changes in the control and treatment group.[21]

The results presented in of Table 2 demonstrate a short term effectiveness of Episode announcements in reducing mean-PM_{10} concentrations across Santiago.

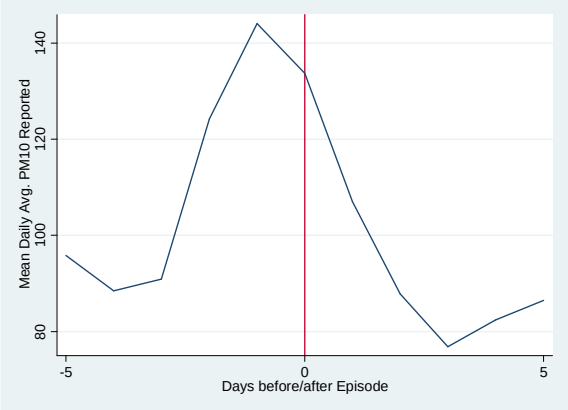
[19] Put another way, the overlapping support requirement of propensity score matching is met in this case by all but one Episode. The unmatched Episode is dropped from the analysis.

[20] It is worth noting that this change is negative for many matched days in the control group suggesting that mean reversion is in fact taking place in PM_{10}concentrations on matched days and the days following matched days.

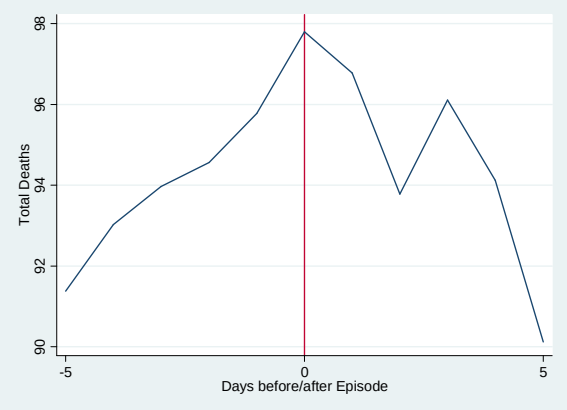
[21] Standard errors were also calculated using a bootstrap approach. See Caliendo and Kopeinig (2008) for a discussion of standard error calculations for propensity score matching approaches.

Figure 5. Event Study of Short-Term Impacts of Episode Announcement

(a) Daily Avg. PM_{10} Concentrations



(b) Avg. Deaths per Day



Notes: Both graphs represent daily, city-wide averages for Santiago grouped by the length of time to the nearest Episode announcement of any level (Alert, Pre-Emergency, and Emergency levels are all included). If a given day is equidistant to two episodes, it is associated with the episode that precedes it. Both graphs show averages based on raw-levels which are not normalized or demeaned in anyway. Pollution data was obtained from Chile's National Ministry of the Environment, and the data on deaths is available from Chile's Ministry of Health.

Table 2. Difference-in-Difference Results for PM_{10}

	Mean Diff-n-Diff
Difference from Day before to Day of Episode	**-28.4177** (6.4631) [6.5887]
Difference from Day -1 to Day 1	**-44.9054** (9.1224) [10.4284]
Difference from Day -1 to Day 2	**-48.1798** (9.4532) [10.0431]
Difference from Day -1 to Day 3	**-33.7685** (9.7056) [10.3522]
Difference from Day -1 to Day 4	**-39.6563** (10.4286) [11.5582]
Difference from Day -1 to Day 5	**-38.6054** (10.5711) [11.4702]

Standard Errors in Parenthesis ( )
Bootstrapped Standard Errors in Square Brackets [ ]

Notes: Results are based on the 36 of 37 Post-PPDA Episodes meeting our separation criteria, which also satisfy the common support restrictions of the Propensity Score Matching approach. All reported values are in terms of PM_{10} concentrations measured in $\mu g/m^3$. Calculations are based on city-wide averages of the daily means of observations for each in-service monitoring station on a given day. The average city-wide PM_{10} level on Episode days was ~$118.9\mu g/m^3$. Pollutant concentration and weather data were obtained from Chile's Ministry of the Environment. Data from 1991-2008 was used for the estimates above. This date restriction was imposed due to weather data availability constraints and changes to the PPDA that were implemented in 2009.
Standard errors computed from propensity score matching methods subject to issues as discussed in Caliendo and Kopeinig (2008).

Bootstrapped standard errors calculated using 500 replications.

We see that, on average, from the time of implementation of the PPDA, through 2008, the announcement of an Environmental Episode led to PM_{10} concentrations that were significantly lower than would otherwise have been expected on the day of the episode and the subsequent two days. Particularly, we see that the day of an Episode experienced average PM_{10} levels that were approximately $30\mu g/m^3$ lower than would have been expected without the Episode announcement. This is a very striking effect given that the mean city-wide level of PM_{10} on Post-PPDA days when Episodes were announce was $118.9\mu g/m^3$. This suggests that an Episode announcement, along with all the government actions and restrictions such an announcement brings into force, lead to immediate PM_{10} concentration reductions of approximately 20% from anticipated levels, with additional air quality benefits continuing for several subsequent days. The magnitude of this day-of estimated effect compares to the high-end of the estimated effects presented by Troncoso et al. (2012), but given that our analysis includes only "stand alone" Episodes, this is to be expected. When we include all Pre-Emergency Episodes in our analysis (as Troncoso et al. (2012) do in their examination), Table A.8 shows a day-of-Episode effect of -15.73 $\mu g/m^3$, or a 9.95% reduction from the day before the Episode. This aligns very closely with the 8.8 to 12.5% reductions estimated for weekday Pre-Emergency Episodes by Troncoso et al. (2012) for the sub-periods of 2000-2006 and 2000-2008, respectively.

In order to assess the short-term effectiveness of Episode announcements on health outcomes, we repeat the previous analysis using daily death count as the outcome variable. Although the population, and therefore number of deaths per day, is growing over time, the Difference-in-Differences approach removes any effects of this trend from the results. Table 3 presents the results for the Difference-in-Differences analysis with daily death counts as the outcome variable.

The results presented in Table 3 are far less striking than those we saw for PM_{10} concentrations, but fit nicely into the existing literature on the effects of short term exposure to elevated levels of PM_{10}. Generally, Table 3 reports estimated difference in difference effects on lives saved by the announcement of an Episode (and all that such an announcement brings along with it) which are small and insignificant. However, the point estimates do have the sign we would expect given that we have shown that Episode announcements lead to improved air quality. We also see that the reductions in deaths persist for several days after the Episode as would be expected given the literature(Schwartz, 2000, Zanobetti et al., 2000). Additionally, although our point estimates are not precisely estimated, their magnitudes are in line with what we would expect given that the mean number of daily deaths in the Metropolitan Region was 95.97 for Post-PPDA days on which Episodes were announced(Zanobetti et al., 2003).

Table 3. Difference-in-Difference Results for Daily Death Count

	Mean Diff-n-Diff
Difference from Day before to Day of Episode	**-1.3234** (3.1695) [3.1666]
Difference from Day -1 to Day 1	**-2.6444** (3.1615) [3.0681]
Difference from Day -1 to Day 2	**-2.7749** (3.0531) [2.9914]
Difference from Day -1 to Day 3	**-4.8079** (2.5291) [2.5526]
Difference from Day -1 to Day 4	**-4.4472** (3.1719) [3.4041]
Difference from Day -1 to Day 5	**-2.0793** (2.7571) [2.9927]

Standard Errors in Parenthesis ( )

Bootstrapped Standard Errors in Square Brackets [ ]

Note: Results are based on the 36 of 37 Post-PPDA Episodes meeting our separation criteria, which also satisfy the common support restrictions of the Propensity Score Matching approach. All values are reported in terms of number of deaths per day in the Metropolitan Region. Data on deaths was obtained from Chile's Ministry of Health. The mean death count on Episode days in the sample is ~96. Similar analyses were run on sub-samples of the deaths data including only those over age-64, and those under age-5, as well as deaths attributed to respiratory and circulatory ailments. Unfortunately, we lack the power to accurately identify any effects in these sub-populations, and not significant results are found. Data from 1992-2008 was used for the estimates above. This date restriction was imposed due to mortality data availability constraints and changes to the PPDA that were implemented in 2009.

Standard errors computed from propensity score matching methods subject to issues as discussed in Caliendo and Kopeinig (2008).

Bootstrapped standard errors calculated using 500 replications.

4.1 Robustness Checks

In all analysis thus-far, we have treated all Post-PPDA Episodes as if they were equivalent and identical, but we saw earlier that some modifications were made to the intensity of driving restriction starting in 2007. To ensure that the inclusion of Post-PPDA Episodes from 2007 and 2008 do not negatively impact our results, we have provided the results of our headline analysis with the years after 2006 omitted. The results from this analysis are presented in Table A.5, which shows no big surprises from the omission of 2007 and 2008 Episodes. The effects on PM_{10} fall slightly, which we would expect given that we dropped Episodes with stricter driving restrictions. Interestingly, deaths seem to climb when we drop the 2007 and 2008 Episodes. This may be due simply to improving medical care or other broader currents which manifest themselves more fully in 2007 and 2008 than in earlier years.

The results reported above rely on average values taken from across the Metropolitan Region for pollution, weather, and death count data. However, it is quite plausible that the impacts of Episode announcements are not spread evenly across the city due to localized topography, weather patterns, emissions profiles, and socio-economic characteristics of the citizenry. In order to assess the possibility of heterogeneous treatment effects by area within the Santiago Metropolitan Region, we repeat the previous analysis based on the pollution and weather data from only one monitoring site at a time. This reanalysis is done for each of the 3 sites for which data is available during both the Pre and Post-PPDA time periods. These three sites are Parque O'Higgins, La Paz/Independencia,[22] and Los Condes. Both the Parque O'Higgins and La Paz stations are located within the topographical basin with most of Santiago. Los Condes however, is a rich suburb that sits in the foothills to the east of and several hundred meters above Santiago's city center. For this reason, air quality is generally much better in Los Condes, and we might expect Episodes to have weaker impacts.

Table A.6 reports the results of the propensity score matched difference-in-differences analyses run on each of our three complete data series individually using area specific PM_{10} and mortality data. Also reported, are full analyses using all comunas within in the Metropolitan Region except that of Los Condes. The results are basically what we would expect given Santiago's topographical characteristics. Pollution levels at Parque O'Higgins (in the heart of urban Santiago) are affected the most by Episode announcement, and Los Condes is affected the least. The estimated effects on deaths generally have the signs we expect, but none are significant. Somewhat interestingly, the point estimates in Los Condes tend to be larger than those in the other comunas.[23] Unfortunately, we do not have statistical power to investigate this any further.

[22]The location of the La Paz monitor was actually changed in 1997 when then MACAM2 network was rolled out, however the new location, known as Independencia is very near the original La Paz location. This analysis therefore treats the observations from the two stations as a single time series.

[23]Note: the higher point estimates for Los Condes do not arise simply from a larger number of daily deaths. Average annual deaths (for the period from 1992-2008) were: 1570.35 for Central Santiago, 841.11 for Independencia, and 1261.89 for Los Condes.

Additionally, the Difference-in-Difference analysis on deaths was run separately for those under the age of 1 and over the age of 64. Results are presented in the first two columns of Table A.7. None of the estimated values are significant at any useful level. Similarly, the analysis was run examining only deaths which were attributed to respiratory ailments and circulatory ailments. Given that these two categories capture most of the ailments for which air pollution is often implicated, we would expect these results to mirror the full mortality results presented in Table 3. Unfortunately, we do not have enough power to draw any conclusions from cause of death specific analyses. The third and fourth columns of Table A.7 nevertheless report the results of these analyses. While none of the estimates are significant at conventional levels, we see the point estimates (and confidence) generally peaking in days 3 and 4 after an Episode, just as we saw in the full analysis.[24]

There may be some question about the validity of our choice to focus only on Episode announcements that "stand alone" in the sense that they are neither preceded or followed by another Episode announcement for 4 days. In order to ensure that the reported results are not being driven by our choice of scope, we have run the analysis described above on all Pre-Emergency and Emergency level Episodes regardless of their proximity to other such Episode announcements. As we might expect, Table A.8 reports the estimated effects of an Episode announcement are reduced by the inclusion of bunched Episode events in the analysis. Since the announcement of Episodes on two subsequent days is indicative that the first Episode did not have the full, desired effect, dropping Episodes that are in close proximity to one another has lead to an overestimation of the ATT. However, even when all Episodes are included, we still have large and significant effects of Episode announcement on PM_{10} levels, and insignificant, but intuitively signed, estimates of the effects on mortality. These results suggest a robustness of our earlier analyses to the use of a broader range of Episodes.

Additionally, we see in both panels of Table A.8 that the effects seem to be growing throughout the reported 6 day time period. This is in contrast to the peaking of effects seen in Tables 2 and 3. The continued increases in effect magnitude due to the inclusion of bunched Episodes can be easily explained. When the analysis includes Episode announcements which are followed closely by other Episodes, some of the effectiveness of the subsequent Episodes is attributed to the first Episode. Thus, effects can appear to be growing, when really what we're seeing is the impact of additional Episode announcements being misattributed. This confounding of effects between closely occurring Episodes is the reason why only stand alone Episodes were analyzed in the central analysis of this paper, thus it is comforting to see that the analysis of the indiscriminate sample leads to the anticipated issues.

Finally, as part of the Propensity Score Matching procedure, we had to make

[24]Additional analyses were run using shares of total deaths from those younger than 1, and older than 64, and shares of deaths attributed to respiratory or circulatory ailments, however the results did not differ significantly from those reported here, and thus were omitted.

some decisions about how closely to require matches to be made. All analysis in this paper up to this point has been run using a caliper of 0.05. Meaning that all Pre-PPDA days with a Propensity Score with 0.05 of the Propensity Score of a Post-PPDA Episode are included in each analysis. It is worthwhile to ensure that our choice of caliper width is not driving our results in any sort of unexpected manner. We again rerun the main analysis of the paper, this time using 3 additional caliper widths (in separate runs), 0.01, 0.1, and 0.15. The results of these 3 runs, along with our original results are presented in Table A.9. The results vary somewhat, as we might expect, but the general relationships and trends remain unchanged regardless of the caliper width chosen.

5 Conclusion

This paper set out to provide a background understanding of the Chilean approach to air pollution in the capital of Santiago, with a special focus on the PPDA which has served as the effective the core of the Chilean approach. Under the PPDA, Chile's air quality policy seeks explicit to improve air quality in Santiago and also reduce the negative health impacts of air pollution in the capital. We show qualitatively that Chile's air pollution policies appear to have lead to large improvements in Santiago's air pollution. Additionally, subsequent to the implementation of Chile's air pollution policies, we observe large reductions in the rates of several types of deaths for which air pollution is often implicated, including deaths among the very young and elderly as well as deaths attributed to respiratory ailments.

Finally, we take a more quantitatively rigorous look at the immediate effects of the implementation of an Environmental Episode. We find the announcement of an Environmental Episode leads to significant reductions in PM_{10} concentrations within Santiago, both on the day of the Episode announcement and on subsequent days. Unfortunately, we do not have the power to precisely identify significant improvements from Episode announcements on deaths, however the point estimates suggest that Episode announcements are associated with small ($<5\%$) reductions in the number of deaths on a given day.

The strategies and approaches employed by the Chilean government since the late 1980s, and in particular under the PPDA, have proved effective at reducing levels of air pollution in both the short and long run, as well as decreasing the frequency of severe air pollution events. Additionally, there is some evidence that health outcomes in the city have improved under the Chilean policies. Thus, though air quality remains a significant issue in Santiago, we find compelling evidence for the overall, to-date, effectiveness of Chile's approach to addressing air pollution in the Santiago Metropolitan Area.

References

D. Bauner and S. Laestadius. The introduction of the automotive catalytic converter in chile. *Journal of Transport Economics and Policy (JTEP)*, 37(2): 157–199, 2003.

M. Caliendo and S. Kopeinig. Some practical guidance for the implementation of propensity score matching. *Journal of economic surveys*, 22(1):31–72, 2008.

R.T. Carson, M.B. Conaway, W.M. Hanemann, R.J. Kopp, K. Martin, RC Mitchell, S. Presser, and PA Ruud. Valuing winter visibility improvement in the grand canyon. In *Association of Environmental and Resource Economists Meeting. New Orleans*, 1992.

K.Y. Chay and M. Greenstone. Does air quality matter? evidence from the housing market. *Journal of Political Economy*, 113(2), 2005.

L.A. Cifuentes, J. Vega, K. Köpfer, and L.B. Lave. Effect of the fine fraction of particulate matter versus the coarse mass and other pollutants on daily mortality in santiago, chile. *Journal of the Air & Waste Management Association*, 50(8):1287–1298, 2000.

L.W. Davis. The effect of driving restrictions on air quality in mexico city. *Journal of Political Economy*, 116(1):38–81, 2008.

DW Dockery and C.A. Pope. Acute respiratory effects of particulate air pollution. *Annual review of public health*, 15(1):107–132, 1994.

D.W. Dockery, C.A. Pope, X. Xu, J.D. Spengler, J.H. Ware, M.E. Fay, B.G. Ferris Jr, and F.E. Speizer. An association between air pollution and mortality in six us cities. *New England journal of medicine*, 329(24):1753–1759, 1993.

F. Dominici, A. McDermott, M. Daniels, S.L. Zeger, and J.M. Samet. Mortality among residents of 90 cities. *Revised analyses of time-series studies of air pollution and health. Special report. Boston, MA: Health Effects Institute*, pages 9–24, 2003.

P.G. Goodman, D.W. Dockery, and L. Clancy. Cause-specific mortality and the extended effects of particulate pollution and temperature exposure. *Environmental Health Perspectives*, 112(2):179, 2004.

E. Gramsch, F. Cereceda-Balic, P. Oyola, and D. Von Baer. Examination of pollution trends in santiago de chile with cluster analysis of pm10 and ozone data. *Atmospheric environment*, 40(28):5464–5475, 2006.

M. Ilabaca, I. Olaeta, E. Campos, J. Villaire, M.M. Tellez-Rojo, and I. Romieu. Association between levels of fine particulate and emergency visits for pneumonia and other respiratory illnesses among children in santiago, chile. *Journal of the Air & Waste Management Association*, 49(9):154–163, 1999.

K Katsouyanni, G Touloumi, E Samoli, A Gryparis, A Le Tertre, Y Monopolis, G Rossi, D Zmirou, F Ballester, A Boumghar, and et al. Confounding and effect modification in the short-term effects of ambient particles on total mortality: Results from 29 european cities within the aphea2 project. *Epidemiology Cambridge Mass*, 12(5):521–531, 2001. URL `http://www.ncbi.nlm.nih.gov/pubmed/11505171`.

James M. Lents, Gerhard Leutart, and Humberto Fuenzalida. Segunda auditoria internacional: Plan de prevencion y descontaminacion atmosferica de la region metropolitana (ppda), March 2006.

G.E. Likens, C.T. Driscoll, and D.C. Buso. Long-term effects of acid rain: Response and recovery of a forest ecosystem. *Science*, 272(5259):244–246, 1996.

C.Y.C. Lin, W. Zhang, and V.I. Umanskaya. The effects of driving restrictions on air quality: São paulo, bogotá, beijing, and tianjin. In *2011 Annual Meeting, July 24-26, 2011, Pittsburgh, Pennsylvania*. Agricultural and Applied Economics Association, 2011.

S. Luechinger. Valuing air quality using the life satisfaction approach*. *The Economic Journal*, 119(536):482–515, 2009.

E. Moretti and M. Neidell. Pollution, health, and avoidance behavior. *Journal of Human resources*, 46(1):154–175, 2011.

M. Neidell. Information, avoidance behavior, and health. *Journal of Human Resources*, 44(2):450–478, 2009.

Chilean Ministry of Health. Supreme decree 32, 1990.

Chilean Ministry of the Environment. Gestiï¿œn de la calidad del aire: Varias dï¿œcadas de esfuerzo, 2007. URL `http://www.votaportuaire.cl/sites/default/files/Gesti%C3%B3n%20de%20la%20calidad%20del%20aire_varias%20d%C3%A9cadas%20de%20esfuerzo.pdf`.

Chilean Ministry of the Secretary General. Supreme decree 46, 2007. URL `http://www.leychile.cl/Navegar?idNorma=260352`.

Ministry of the Secretary General. Establishment of the plan of prevention and atmospheric decontamination of the metropolitan region. http://www.votaportuaire.cl/sites/default/files/D.S.January 1998.

World Health Organization. Air quality and health, September 2011.

B. Ostro, J.M. Sanchez, C. Aranda, G.S. Eskeland, et al. Air pollution and mortality: results from a study of santiago, chile. *Journal of Exposure Analysis and Environmental Epidemiology*, 6(1):97, 1996.

B.D. Ostro. The effects of air pollution on work loss and morbidity. *Journal of Environmental Economics and Management*, 10(4):371–382, 1983.

B.D. Ostro, G.S. Eskeland, J.M. Sanchez, and T. Feyzioglu. Air pollution and health effects: A study of medical visits among children in santiago, chile. *Environmental Health Perspectives*, 107(1):69, 1999.

P. Pino, T. Walter, M. Oyarzun, R. Villegas, and I. Romieu. Fine particulate matter and wheezing illnesses in the first year of life. *Epidemiology*, 15(6):702, 2004.

C.A. Pope, M.J. Thun, M.M. Namboodiri, D.W. Dockery, J.S. Evans, F.E. Speizer, C.W. Heath Jr, et al. Particulate air pollution as a predictor of mortality in a prospective study of us adults. *American journal of respiratory and critical care medicine*, 151(3):669–674, 1995.

C.A. Pope, R.T. Burnett, M.J. Thun, E.E. Calle, D. Krewski, K. Ito, and G.D. Thurston. Lung cancer, cardiopulmonary mortality, and long-term exposure to fine particulate air pollution. *JAMA: the journal of the American Medical Association*, 287(9):1132–1141, 2002.

C.A. Pope, R.T. Burnett, G.D. Thurston, M.J. Thun, E.E. Calle, D. Krewski, and J.J. Godleski. Cardiovascular mortality and long-term exposure to particulate air pollution epidemiological evidence of general pathophysiological pathways of disease. *Circulation*, 109(1):71–77, 2004.

Margarita Prendez, Gerardo Alvarado, and Italo Serey. *Air Quality Monitoring, Assessment and Management*, chapter Some Guidelines to Improve the Air Quality Management of Santiago, Chile. InTech, http://www.intechopen.com/books/air-quality-monitoring-assessment-and-management/some-guidelines-to-improve-the-air-quality-management-of-santiago-chile, 2011.

J. Rutllant and R. Garreaud. Meteorological air pollution potential for santiago, chile: towards an objective episode forecasting. *Environmental Monitoring and Assessment*, 34(3):223–244, 1995.

P.E. Saide, G.R. Carmichael, S.N. Spak, G. Laura, A.E. Osses, M.A. Mena-Carrasco, and P. Mariusz. Forecasting urban pm10 and pm2. 5 pollution episodes in very stable nocturnal conditions and complex terrain using wrf-chem co tracer model. *Atmospheric Environment*, 2011.

P.H.N. Saldiva, C.A. Pope III, J. Schwartz, D.W. Dockery, A.J. Lichtenfels, J.M. Salge, I. Barone, and G.M. Bohm. Air pollution and mortality in elderly people: a time-series study in sao paulo, brazil. *Archives of Environmental Health: An International Journal*, 50(2):159–163, 1995.

M. Salinas, J. Vega, et al. The effect of outdoor air pollution on mortality risk: an ecological study from santiago, chile. *World health statistics quarterly. Rapport trimestriel de statistiques sanitaires mondiales*, 48(2):118, 1995.

P. Sanhueza H, C. Vargas R, and J. Jiménez P. Mortalidad diaria en santiago y su relación con la contaminación del aire; daily mortality in santiago and its relationship with air pollution. *Rev. méd. Chile*, 127(2):235–42, 1999.

J. Schreifels. Emissions trading in santiago, chile: A review of the emission offset program of supreme decree no. 4. 2011.

J. Schwartz. The distributed lag between air pollution and daily deaths. *Epidemiology*, 11(3):320, 2000.

R. Troncoso, L. de Grange, and L. Cifuentes. Effects of environmental alerts and pre-emergencies on pollutant concentrations in santiago, chile. *Atmospheric Environment*, 2012.

B. Viard and S. Fu. The effect of beijing's driving restrictions on pollution and economic activity. *MPRA Paper*, 2011.

A. Zanobetti, J. Schwartz, and D.W. Dockery. Airborne particles are a risk factor for hospital admissions for heart and lung disease. *Environmental health perspectives*, 108(11):1071, 2000.

A. Zanobetti, J. Schwartz, E. Samoli, A. Gryparis, G. Touloumi, J. Peacock, R.H. Anderson, A. Le Tertre, J. Bobros, M. Celko, et al. The temporal pattern of respiratory and heart disease mortality in response to air pollution. *Environmental health perspectives*, 111(9):1188, 2003.

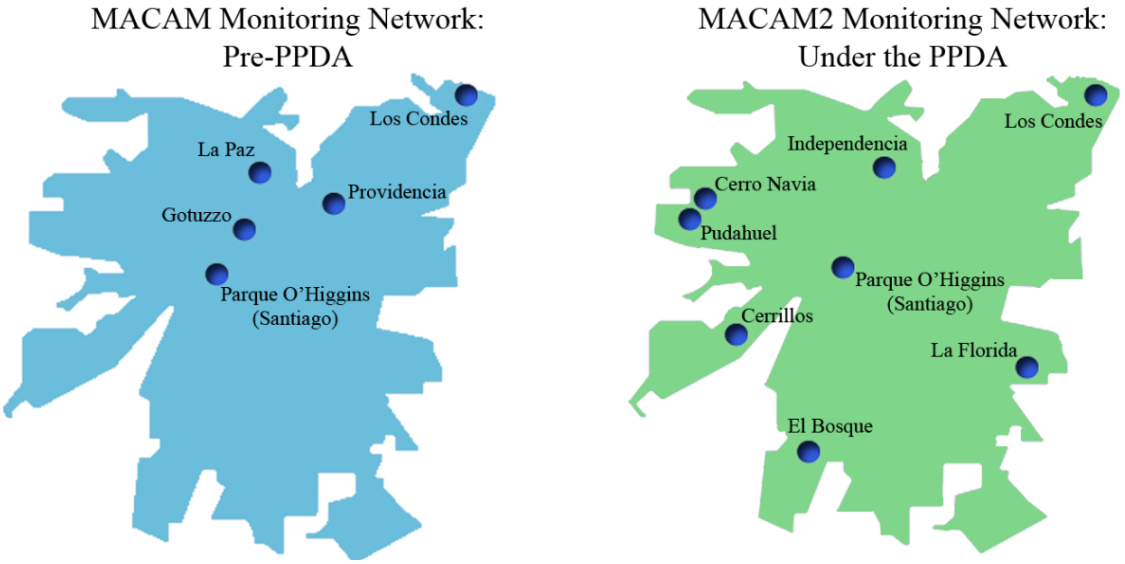
J.G. Zivin and M. Neidell. The impact of pollution on worker productivity. *NBER Working Paper*, 2011.

Appendix

Table A.1. Original PPDA Protocols

Episode Level	Protocols
Seasonal:	
April-August	• Restricted weekday usage of 20% of vehicles without catalytic converters • A citywide traffic plan will be implemented to minimize the effects of the vehicular restrictions
Episodic:	
Alert	• Restricted usage of 40% (weekdays) or 20% (weekends) of vehicles without catalytic converters • Prohibition on the use of uncertified residential wood or biomass heating units
Pre-Emergency	• Restricted usage of 60% (weekdays) or 40% (weekends) of vehicles without catalytic converters • Restricted usage of 20% (all days) of vehicles with catalytic converters • Shut down of stationary emitters without proven emission levels $< 32mg/m^3N$ • Physical Education classes and community sports activities may be suspended by the Ministry of Education • Implementation of more intensive traffic and public transportation plan • Increased enforcement of restrictions on mobile and stationary sources of air pollution • Increased and focused street sweeping and cleaning activities • Increased Metro service schedule implemented • Prohibition of the use of wood or other biomass for residential heating
Emergency	• Restricted usage of 80% (weekday) or 60% (weekend) of vehicles without catalytic converters • Restricted usage of 40% (all days) of vehicles with catalytic converters • Shut down of stationary emitters without proven emission levels $< 28mg/m^3N$ • Physical Education classes and community sports activities may be suspended by the Ministry of Education • Implementation of more intensive traffic and public transportation plan • Increased enforcement of vehicle usage restrictions • Increased and focused street sweeping and cleaning activities • Increased Metro service schedule implemented • Prohibition of the use of wood or other biomass for residential heating

Figure A.1. Maps of Santiago with Pollution Monitoring Networks



Maps are an adaptions of those produced in of the Environment (2007).

Figure A.2. PM_{10} Concentrations by Station: 1988-2010

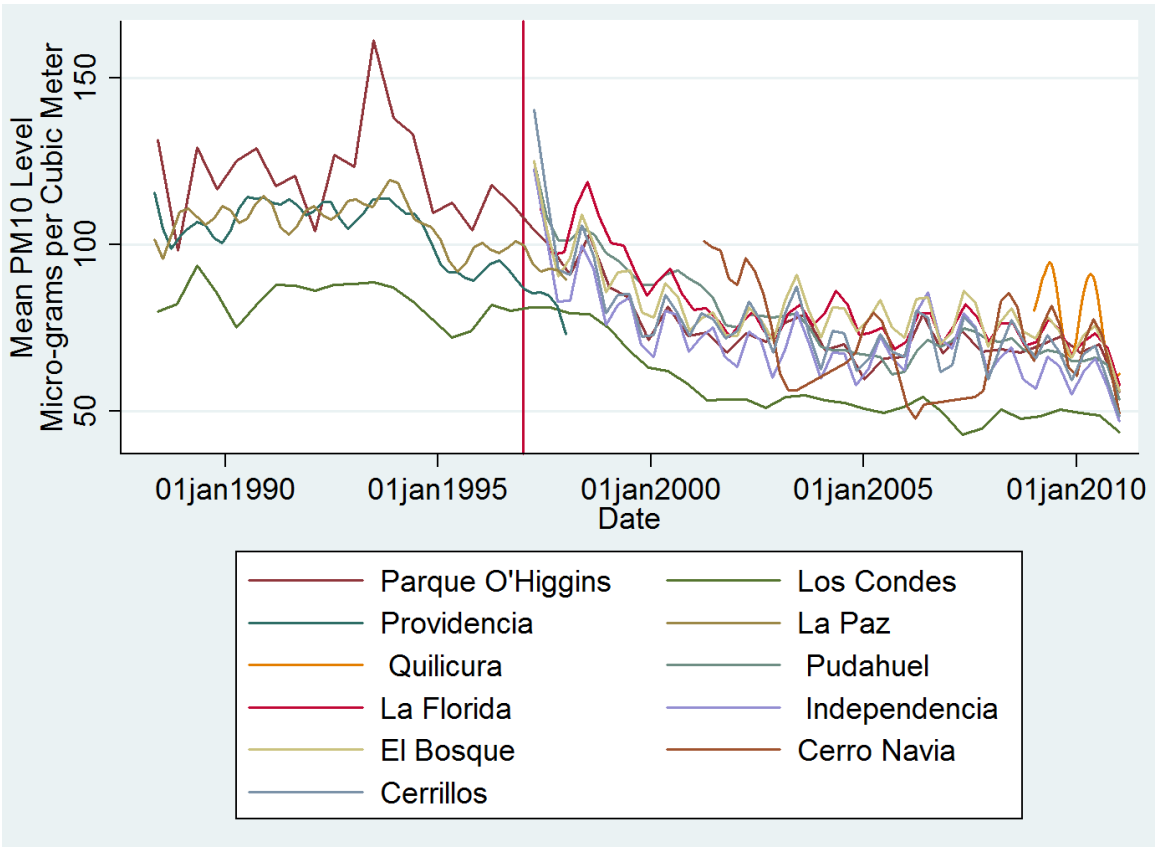


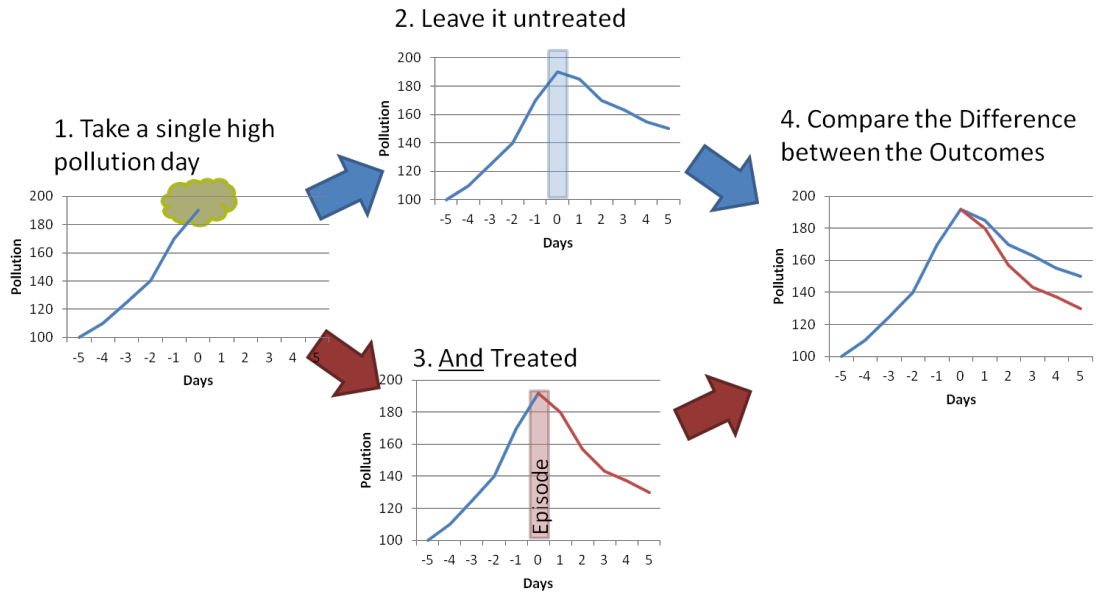
Table A.2. Metropolitan Region Population

Year	Total	Over Age 64	Under Age 5
1990	5,190,548	310,543	570,408
1991	5,292,913	321,587	572,497
1992	5,395,325	332,631	574,587
1993	5,497,718	343,675	576,676
1994	5,600,153	354,719	578,766
1995	5,702,576	365,763	580,855
1996	5,796,305	378,536	569,003
1997	5,890,048	391,309	557,151
1998	5,983,768	404,083	545,298
1999	6,077,530	416,856	533,446
2000	6,171,283	429,629	521,594
2001	6,244,780	444,255	518,102
2002	6,318,299	458,881	514,610
2003	6,391,827	473,506	511,117
2004	6,465,348	488,132	507,625
2005	6,538,896	502,758	504,133
2006	6,607,805	523,035	504,696
2007	6,676,745	543,311	505,259
2008	6,745,651	563,588	505,823
2009	6,814,630	583,864	506,386
2010	6,883,563	604,141	506,949

Notes: Data available from Chile's National Institute of Statistics. All values represent National Institute of Statistic's population projections based on the 2002 National Census.

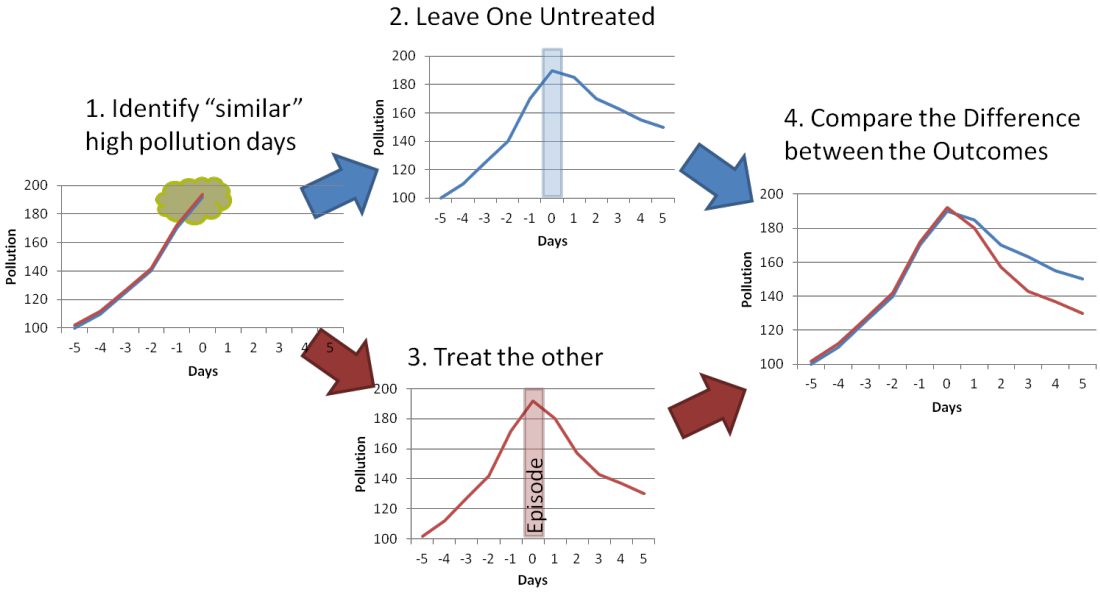
Figure A.3. Empirical Method

Empirical Ideal



Empirical Reality

Unfortunately, we cannot observe the counterfactuals of each Episode day, therefore we must look for an appropriate comparison



Since our outcome of interest is a change, the comparison of differences becomes a Difference-in-Differences. We now need only to identify "similar" high pollution days for comparison.

Table A.3. Probit Results from PM_{10} matching exercise

VARIABLES	Coeff. (Std. Error)	VARIABLES (cont'd)	Coeff. (Std. Error)
Prior Day's mean PM_{10}	0.00433	3 Days Prior Mean Temp Squared	-0.0988
	(0.00338)		(0.0761)
2 Days Prior mean PM_{10}	-0.00123	3 Days Prior Mean Temp Cubed	0.000626
	(0.00318)		(0.000496)
Prior Day's Mean Temp	0.737	4 Days Prior Mean Temp Squared	0.136**
	(2.970)		(0.0645)
2 Days Prior mean Temp	1.222	4 Days Prior Mean Temp Cubed	-0.000926**
	(2.264)		(0.000431)
3 Days Prior mean Temp	5.098	Sunday Dummy	0.310
	(3.870)		(0.353)
4 Days Prior mean Temp	-6.518**	Monday Dummy	0.366
	(3.195)		(0.367)
Prior Day's Mean Wind	-0.324**	Wednesday Dummy	0.0902
	(0.140)		(0.339)
2 Days Prior mean Wind	-0.514***	Thursday Dummy	0.101
	(0.150)		(0.352)
3 Days Prior mean Wind	0.0127	Friday Dummy	0.245
	(0.0151)		(0.345)
4 Days Prior mean Wind	0.0912	April Dummy	-0.768
	(0.0979)		(1.101)
Prior Day's Mean Dewpoint	-0.107***	May Dummy	-1.477
	(0.0251)		(1.234)
2 Days Prior Mean Dewpoint	-0.000576	June Dummy	-1.326
	(0.00513)		(1.255)
3 Days Prior Mean Dewpoint	-0.000605	July Dummy	-0.689
	(0.00270)		(1.254)
4 Days Prior Mean Dewpoint	-0.0384	August Dummy	-1.100
	(0.0253)		(1.262)
Prior Day's Mean Temp Squared	-0.00908	September Dummy	-1.439
	(0.0573)		(1.269)
Prior Day's Mean Temp Cubed	3.12e-05	Constant	-15.74
	(0.000365)		(63.92)
2 Days Prior Mean Temp Squared	-0.0257		
	(0.0438)	Observations	955
2 Days Prior Mean Temp Cubed	0.000186	Pseudo R-Square	0.323
	(0.000281)		

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table A.4. Table of Annual Means

	Pre-PPDA	Post-PPDA	Change
Annual Death Count	28,189	31,334	11.2%
Over 64 Death Count	17,430	20,983	20.4%
Under 1 Death Count	1,244	818	-34.3%
Under 5 Death Count	1,458	958	-34.3%
Death Count due to Respiratory Ailments	3,726	3,366	-9.7%
Death Count due to Circulatory Ailments	7,948	8,735	9.9%
Death Count due to Cancer	6,508	7,368	13.2%
Death Count due to Accidents	2,917	2,057	-29.5%
Average Age of Deaths	64.00	67.74	5.8%
Average CO Level (ppm)	2.46	1.03	-58.2%
Average NO Level (ppb)	56.76	46.27	-18.5%
Average NO_2 Level (ppb)	32.28	22.57	-30.1%
Average $Ozone$ Level (ppb)	20.43	15.99	-21.7%
Average $PM_{2.5}$ Level ($\mu g/m^3$)	53.69	33.23	-38.1%
AveragePM_{10} Level ($\mu g/m^3$)	104.21	76.15	-26.9%
Average Temperature	57.85	58.66	1.4%
Average Daily Max Temp	74.30	74.35	0.1%
Average Daily Min Temp	44.68	45.83	2.6%
Average Windspeed	4.92	4.62	-6.2%
Average Dew-point	56.17	45.72	-18.6%
Average Episode (any level)	7.67	25.92	238.0%
Average Stand Alone Episode	-	3.083	-
Average Population	5,547,498	6,342,665	14.3%
Male Population	2,693,118	3,091,522	14.8%
Female Population	2,854,380	3,251,143	13.9%
Over 64 Years Old	349,485	469,945	34.5%
Under 5 Years Old	575,397	519,071	-9.8%
Death Rate	0.5035%	0.4939%	-1.9%
Over 64 Death Rate	4.9106%	4.4869%	-8.6%
Under 4 Death Rate	0.2532%	0.1839%	-27.4%
Death Rate due to Respiratory Ailments	0.0653%	0.0534%	-18.3%
Death Rate due to Circulatory Ailments	0.1395%	0.1376%	-1.3%
Death Rate due to Cancer	0.1142%	0.1160%	1.6%
Death Rate due to Accidents	0.0512%	0.0321%	-37.3%
Day Count in Period	2,192	4,383	-

Note: All values except "Day Count in Period" are annual averages.

Table A.5. Propensity Score D-n-D with 2007 and 2008 Omitted

	PM_{10}		Deaths	
	1991-2008	1991-2006	1992-2008	1992-2006
	Mean Diff-n-Diff	Mean Diff-n-Diff	Mean Diff-n-Diff	Mean Diff-n-Diff
Difference from before to day of Episode	**-28.4177**	**-25.1822**	**-1.3234**	**-2.4368**
	(6.4631)	(6.7500)	(3.1695)	(3.2216)
Difference from Day -1 to day 1	**-44.9054**	**-40.5017**	**-2.6444**	**-7.1081**
	(9.1224)	(10.1412)	(3.1615)	(3.0831)
Difference from Day -1 to day 2	**-48.1798**	**-44.6162**	**-2.7749**	**-2.5796**
	(9.4532)	(10.2603)	(3.0531)	(3.3648)
Difference from Day -1 to day 3	**-33.7685**	**-28.2557**	**-4.8079**	**-7.0531**
	(9.7056)	(10.2508)	(2.5291)	(2.4219)
Difference from Day -1 to day 4	**-39.6563**	**-30.0027**	**-4.4472**	**-6.4972**
	(10.4286)	(11.0759)	(3.1719)	(3.4102)
Difference from Day -1 to day 5	**-38.6054**	**-33.2000**	**-2.0793**	**-2.8890**
	(10.5711)	(11.5794)	(2.7571)	(2.8042)

Notes: The first column in each section of the table above represents a reproduction of the results presented in Tables 2 & 3 above. This is done for comparison sake, as subsequent columns report the results of the same analyses being repeated on subsets of the original data. The PM_{10} analysis (left half of table) uses data beginning in 1991 due to weather data availability constraints. Similarly, the mortality analysis (right half of table) uses data only from 1992 onward due to mortality data availability.

Table A.6. Station Specific PM_{10}: Robustness

	PM_{10}					Deaths				
	Metro Region (from Above)	Parque O'Higgins	La Paz	Los Condes	All Monitors but Los Condes	Metro Region (from Above)	Parque O'Higgins	La Paz	Los Condes	All Monitors but Los Condes
Difference from Day before to Day of Episode	**-28.4177**	**-27.5609**	**-25.1666**	**-22.4534**	**-28.1711**	**-1.3234**	**-0.0879**	**0.0183**	**-0.0851**	**-0.4884**
	(6.4631)	(11.0519)	(7.2117)	(6.2965)	(8.1763)	(3.1695)	(0.6538)	(0.4620)	(0.5676)	(1.0024)
Difference from Day -1 to Day 1	**-44.9054**	**-44.3635**	**-33.3965**	**-32.0632**	**-42.3035**	**-2.6444**	**0.1637**	**-0.1805**	**0.5808**	**-0.4607**
	(9.1224)	(14.6646)	(9.6721)	(9.2633)	(11.0142)	(3.1615)	(0.5983)	(0.4043)	(0.6438)	(0.8653)
Difference from Day -1 to Day 2	**-48.1798**	**-58.4587**	**-39.8624**	**-23.4699**	**-52.3890**	**-2.7749**	**-0.3876**	**0.4655**	**-0.9688**	**0.0245**
	(9.4532)	(14.7923)	(9.8181)	(9.8138)	(11.2550)	(3.0531)	(0.5903)	(0.5064)	(0.5712)	(1.0132)
Difference from Day -1 to Day 3	**-33.7685**	**-39.0973**	**-31.3531**	**-20.1577**	**-38.6553**	**-4.8079**	**-0.7272**	**-0.5055**	**-0.0889**	**-2.1577**
	(9.7056)	(15.2251)	(9.9633)	(8.4730)	(11.6339)	(2.5291)	(0.6168)	(0.4807)	(0.6186)	(0.9094)
Difference from Day -1 to Day 4	**-39.6563**	**-40.4754**	**-39.6844**	**-20.5057**	**-45.1039**	**-4.4472**	**-0.9562**	**-0.4544**	**0.8027**	**-1.7698**
	(10.4286)	(16.2354)	(11.4144)	(8.8608)	(12.8903)	(3.1719)	(0.4738)	(0.4124)	(0.6075)	(0.7634)
Difference from Day -1 to Day 5	**-38.6054**	**-40.3916**	**-37.7530**	**-26.9134**	**-45.6081**	**-2.0793**	**0.0018**	**-0.7327**	**0.4864**	**-1.2383**
	(10.5711)	(16.8249)	(11.2250)	(8.7242)	(13.0557)	(2.7571)	(0.6313)	(0.4424)	(0.4570)	(0.9579)
Episodes on Support	35	33	35	34	35	35	34	36	36	35
Episodes off Support	1	2	1	1	1	1	2	0	0	1

Notes: The first column in each section of the table above represents a reproduction of the results presented in Tables 2 & 3 above. This is done for comparison sake, as subsequent columns report the results of the same analyses being repeated on subsets of the original data. The PM_{10} analysis (left half of table) uses data from 1991-2008 due to data availability. Similarly, the mortality analysis (right half of table) uses data only from 1992-2008 due to mortality data availability.

Table A.7. Difference-in-Differences Results by Sub-Population and Cause of Death

	Over 64	Under 1	Respiratory Ailments	Circulatory Ailments
Difference from Day before to Day of Episode	**-0.2089**	**0.7246**	**0.2572**	**0.4779**
	(2.4221)	(0.4682)	(0.9595)	(1.3555)
Difference from Day -1 to Day 1	**-1.6020**	**0.5096**	**0.3787**	**-0.2094**
	(2.2681)	(0.5173)	(1.0036)	(1.3861)
Difference from Day -1 to Day 2	**-1.2015**	**0.0146**	**-0.5053**	**-0.5840**
	(2.2196)	(0.5898)	(1.1264)	(1.4190)
Difference from Day -1 to Day 3	**-4.5369**	**0.0331**	**-1.0904**	**-1.1087**
	(1.8959)	(0.4865)	(0.9461)	(1.0873)
Difference from Day -1 to Day 4	**-3.5006**	**0.1943**	**-0.3076**	**-1.6001**
	(2.4280)	(0.6073)	(0.9418)	(1.6206)
Difference from Day -1 to Day 5	**-1.0880**	**-0.5666**	**-1.0994**	**-0.3182**
	(2.0661)	(0.4835)	(0.9606)	(1.3864)

Notes: All estimates are based on levels of sub-population or cause-specific deaths. Age-range sub-population mortality statistics are based on age of deceased as reported in the Chilean Mortality statistics available from the Chilean Ministry of Health. Similarly, cause-specific deaths are identified from the Ministry of Health data using the coding based on the World Health Organization's"International Statistical Classification of Diseases and Related Health Problems or "ICD" codes. For the years up to and including 1996, Chile reported deaths using the ICD-9 convention, while in the years from 1997 onward, ICD-10 conventions are used. In the table above, deaths attributed to Respiratory Ailments were those coded with ICD-9 values greater than 459 and less than 520, and ICD-10 codes with the "J" prefix. Similarly, deaths attributed to Circulatory Ailments were those with ICD-9 codes from 390 to 459 or ICD-10 codes beginning with "I".

Table A.8. Difference-in-Differences Results using All Post-PPDA Episode Announcements as Treatment

	PM_{10}	Deaths
	Mean Diff-n-Diff	**Mean Diff-n-Diff**
Difference from Day before to Day of Episode	**-15.7357**	**-0.2151**
	(5.3496)	(2.0475)
Difference from Day -1 to Day 1	**-24.6188**	**-1.1868**
	(7.2807)	(2.0520)
Difference from Day -1 to Day 2	**-27.8477**	**-2.7142**
	(7.6254)	(1.9796)
Difference from Day -1 to Day 3	**-22.3757**	**-3.8398**
	(8.0578)	(1.9249)
Difference from Day -1 to Day 4	**-23.6688**	**-3.9883**
	(8.3668)	(2.1690)
Difference from Day -1 to Day 5	**-27.9185**	**-4.4713**
	(8.2876)	(1.9974)
Pseudo R-Squared	**0.3534**	**0.3534**
Episodes on Support	91	91
Episodes off Support	0	0

Note: This approach provides a larger sample for analysis. In both the PM_{10} and Deaths analyses, all 91 Post-PPDA Episodes are matched to (one or more) Pre-PPDA days under our chosen radius matching approach, and are thus used in these analysis.

Table A.9. Caliper Size Robustness Check

Dependent Variable	**PM10: Mean Diff-n-Diff**				**Deaths: Mean Diff-n-Diff**			
Caliper Size	**0.01**	**0.05**	**0.1**	**0.15**	**0.01**	**0.05**	**0.1**	**0.15**
Difference from before to day of Episode	-23.1053	-28.4177	-28.3909	-29.5040	0.7996	-1.3234	-1.5911	-1.5506
	(7.6679)	(6.4631)	(6.1519)	(5.8294)	(3.3647)	(3.1695)	(3.1127)	(3.0556)
Difference from Day -1 to day 1	-49.2111	-44.9054	-44.1182	-45.9460	-3.3336	-2.6444	-2.6977	-2.7522
	(10.7257)	(9.1224)	(8.7786)	(8.4265)	(3.4836)	(3.1615)	(3.1052)	(3.0486)
Difference from Day -1 to day 2	-50.2671	-48.1798	-51.7671	-55.0646	-1.2958	-2.7749	-2.7446	-2.7973
	(11.3223)	(9.4532)	(9.0588)	(8.6530)	(3.3876)	(3.0531)	(2.9855)	(2.9173)
Difference from Day -1 to day 3	-36.9653	-33.7685	-38.6718	-43.9149	-3.8134	-4.8079	-4.5458	-4.5340
	(11.5724)	(9.7056)	(9.2932)	(8.8685)	(2.8146)	(2.5291)	(2.4446)	(2.3582)
Difference from Day -1 to day 4	-45.1104	-39.6563	-43.1033	-46.6380	-4.4068	-4.4472	-4.7773	-5.3286
	(12.1821)	(10.4286)	(10.0212)	(9.6034)	(3.4273)	(3.1719)	(3.1042)	(3.0360)
Difference from Day -1 to day 5	-41.7753	-38.6054	-42.3264	-45.4744	-0.4868	-2.0793	-2.7310	-2.9738
	(12.4305)	(10.5711)	(10.2130)	(9.8478)	(2.9254)	(2.7571)	(2.6800)	(2.6018)
On Support	32	35	35	35	32	35	35	35
Off Support	4	1	1	1	4	1	1	1

Notes: The 2nd column in each section of the table above represents a reproduction of the results presented in Tables 2 & 3 above, it is boxed to mark this fact. These columns are included for comparison sake, as the other columns in the table report the results of the similar analyses being repeated on the same data, changing only the caliper size used for the Propensity Score Matching Procedure. The PM_{10} analysis (left half of table) uses data from 1991-2008 and the mortality analysis (right half of table) uses data only from 1992-2008 due to data availability.