8-1-2012

Essays on Experimental and Quasi-Experimental Policy Design and Evaluation

Juan Jose Miranda Montero

PERMISSION TO BORROW

NOTICE TO BORROWERS

All dissertations deposited in the Georgia State University Library must be used only in accordance with the stipulations prescribed by the author in the preceding statement.

The author of this dissertation is:
Juan José Miranda Montero
5480 Wisconsin Ave, Apartment 811
Chevy Chase, MD 20815

The director of this dissertation is:
Dr. Paul J. Ferraro
Andrew Young School of Policy Studies
Georgia State University
P. O. Box 3992
Atlanta, GA 30302-3992

Users of this dissertation not regularly enrolled as students at Georgia State University are required to attest acceptance of the preceding stipulations by signing below. Libraries borrowing this dissertation for the use of their patrons are required to see that each user records here the information requested.

			Type of use
Name of User	Address	Date	(Examination only or copying)

ESSAYS ON EXPERIMENTAL AND QUASI-EXPERIMENTAL POLICY DESIGN AND EVALUATION

BY

JUAN JOSE MIRANDA MONTERO

A Dissertation Submitted in Partial Fulfillment
of the Requirements for the Degree
of
Doctor of Philosophy
in the
Andrew Young School of Policy Studies
of
Georgia State University

GEORGIA STATE UNIVERSITY
2012

ACCEPTANCE

This dissertation was prepared under the direction of the candidate's Dissertation Committee. It has been approved and accepted by all members of that committee, and it has been accepted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Economics in the Andrew Young School of Policy Studies of Georgia State University.

Dissertation Chair:	Dr. Paul J. Ferraro
Committee:	Dr. H. Spencer Banzhaf
	Dr. Craig McIntosh
	Dr. Rusty Tchernis

Electronic Version Approved:
Mary Beth Walker, Dean
Andrew Young School of Policy Studies
Georgia State University
July 2012

ACKNOWLEDGMENTS

This dissertation was made possible by the support of many people. I owe a debt of gratitude to all of the friends, family members, and teachers who have facilitated my personal and academic growth over the years. All of you know who you are.

First and foremost, I would like to thank to Dr. Paul Ferraro for his great mentorship, encouragement, support, and friendship. He has been instrumental in the achievement of my Ph.D. His influence on my professional life has been far greater than that of a dissertation chair. Second, I would like to thank to Dr. Roxana Barrantes, my Peruvian mentor, former boss, and friend, for her unconditional support, both personally and academically. I am also indebted to the members of my dissertation committee, Dr. Spencer H. Banzhaf, Dr. Craig McIntosh, and Dr. Rusty Tchernis, for their insightful comments and suggestions.

I would like to dedicate this dissertation to Lorena, my wife, with love and gratitude for her patience and encouragement. I am grateful for her support at all times and everywhere. To Jaime and Juana, my parents, their dedication and hard work have been my examples. To Jaime, my brother, for all that he has done for me. To Andrea and Akemi, my nieces, who brought happiness and light to my family.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF APPENDICES

ABSTRACT

ESSAYS ON EXPERIMENTAL AND QUASI-EXPERIMENTAL POLICY DESIGN AND EVALUATION

BY

JUAN JOSE MIRANDA MONTERO

July 2012

Committee Chair: Dr. Paul J. Ferraro

Major Department: Economics

This dissertation comprises two essays. The unifying theme is the evaluation of non-pecuniary (information or norm based messages) conservation programs. These types of policies are widely applied in developing and developed countries to promote conservation, however, their empirical evidence and their effectiveness are not well documented. Each chapter examines some methodological facets of the heterogeneity of non-pecuniary conservation programs and the reliability of non-experimental methods (program evaluation and econometric techniques) to evaluate treatment effects in the context of non-pecuniary conservation programs.

Chapter I evaluates causal channels, heterogeneous responses and policy impacts of information-based programs. Policymakers often rely on non-pecuniary, information-based programs to achieve social objectives. Using data from a water conservation information campaign implemented as a randomized controlled trial, we evaluate heterogeneous household responses. Understanding such heterogeneity is important for improving the cost-effectiveness of non-pecuniary programs and extending them to other populations. We find little evidence of heterogeneous responses to purely technical information or traditional conservation messages, but strong evidence of heterogeneous responses to pro-social messages that highlight social norms: wealthier, owner-occupied households, and households that use more water are more responsive. In contrast, these subgroups tend to be least responsive to pecuniary incentives. Combining theory and data, we also shed light on the mechanisms through which the treatment effects arise: norm-based messages induce behavioral (variable-cost), rather than technological (fixed-cost), changes in outdoor water use and work through social preferences, rather than by serving as signals of privately efficient behavior to boundedly rational agents.

Chapter II assesses the performance of non-experimental evaluation designs. In environmental policy, as in other areas of social policy, randomized evaluation designs are difficult to implement and thus researchers must rely on non-experimental empirical designs to evaluate program impacts. Yet there is considerable debate about whether non-experimental designs can generate accurate estimates of program impact. Design-replication studies assess the ability of non-experimental designs to replicate unbiased (experimental) estimators of program impact. Our design-replication study uses, as a benchmark, a large-scale randomized field experiment that tested the effectiveness of messages designed to induce voluntary reductions in water consumption during a drought. We find that by following best practices described in the literature, panel data methods combined with matching methods, using a rich covariate set including baseline outcomes, can replicate the experimental estimates. However, minor deviations from these practices or the use of fewer covariates can yield grossly inaccurate estimates of treatment effects. In particular, we find that fixed-effects panel data methods fail to replicate experimental estimates unless matching methods are used to pre-process the data.

CHAPTER I:
HETEROGENEOUS TREATMENT EFFECTS AND MECHANISMS IN INFORMATION-BASED ENVIRONMENTAL POLICIES: EVIDENCE FROM A LARGE-SCALE NATURAL FIELD EXPERIMENT

1. Introduction

Non-pecuniary, information-based environmental policy strategies have long been used to influence individual decision-making (e.g., Smith et al. 1990; Smith and Desvousges 1990) and are growing in popularity in all social policy fields (Thaler and Sunstein 2008; House of Lords 2011). Such strategies include norm-based persuasive messages, commitment devices, changes to default options and the provision of technical information to lower transaction costs of information acquisition. Under standard economic assumptions of perfectly-informed, rational, self-interested agents, these strategies should be ineffective. However, under behavioral theories that include other-regarding preferences or bounded rationality, they may be effective. A growing empirical literature in economics and psychology suggests that such strategies can indeed affect policy-relevant behaviors (e.g., Bui and Meyer 2003; Duflo and Saez 2003; Jin and Leslie 2003; Bjørner et al. 2004; Schultz et al. 2007; Goldstein et al. 2008; Bennear and Olmstead 2008; Allcott and Mullainathan 2010; Habyarimana and Jack 2011).

In the context of environmental policies and programs, the conceptual and empirical foundations of such strategies remain under-researched (Shogren and Taylor 2008). A new literature uses randomized controlled trials, which are rare in environmental economics (Greenstone and Gayer 2009; List and Price, forthcoming), to test the impacts of non-pecuniary, norm-based messages on environmental outcomes such as energy use (e.g. Ayres, Raseman and Shih 2009; Yoeli 2009; Allcott 2010; Costa and Kahn 2010) and water use (e.g. Ferraro and Price forthcoming; Ferraro, Miranda and Price 2011). These studies find that sending pro-social messages and social comparisons that contrast own consumption to peer-group consumption can reduce, on average, water and energy consumption. However, a more important question is to understand the mechanisms through which the messages affect behavior, and also it is important to move forward from mean effects to understand the variability of results and recognize the types of households that are most responsive to these types of messages.

Studying heterogeneous treatment effects yields policy and research-relevant insights (Heckman, Smith and Clements 1997; Angrist 2004; Djebbari and Smith 2008). For example, it can help policy makers more cost-effectively target the treatments to subgroups that are most responsive. It can also help strengthen the external validity of randomized controlled trials. The mean effects of the same experimental design could be different when applied in other populations with different distributions of observable characteristics (Hotz et al. 2005). Third, by combining theory and subgroup analysis, one can explore potential mechanisms through which the causal effects are generated. As Deaton (2010) noted in his critique of the way in which randomized controlled trials are done in economics, we need to move beyond determining *whether* a treatment is effective to determining *why* it is effective.[1]

We study heterogeneity and mechanisms in the context of the large-scale field experiment. The experiment was run in 2007 in partnership with the Cobb County Water System in metropolitan Atlanta, Georgia, USA. To induce voluntary reductions in water use during a drought, three types of messages were sent, at random, to households. The three treatments comprise: (i) a tip sheet with information about reducing water consumption (pure information message), (ii) a tip sheet and a personalized letter promoting pro-social behavior (weak social norm message); and (iii) a tip sheet, personalized letter promoting pro-social behavior, and a social comparison of the household's water consumption with the median county consumption (strong social norm message). Each treatment group comprised roughly 11,700 houses and the control group comprised roughly 71,800 houses. Ferraro and Price (forthcoming) report short-term average treatment effects and Ferraro, Miranda and Price (2011) extend the analysis to report longer-term average treatment effects.

When estimating heterogeneous treatment effects from experimental or non-experimental data, there is a substantial risk of labeling spurious correlations as conditional treatment effects. We mitigate this risk through our experimental design and a multi-step framework that trades increasingly stringent assumptions for increasingly precise characterizations of the heterogeneity. We find little evidence of heterogeneous responses to the pure information and weak social norm messages, but strong evidence of heterogeneous responses to the strong social norm message:

[1] Evidence of heterogeneous responses also informs future observational studies that may use instrumental variables to estimate treatment effects from information-based strategies. Heterogeneity implies that the estimates should be interpreted as local average treatment effects (LATE) rather than population average treatment effects (ATE).

wealthier households, owner-occupied households, and households that use more water are more responsive. Interestingly, these are among the households identified in the literature to be *least* responsive to pecuniary incentives. Also, in contrast to predictions from the psychology literature that low resource users may respond to the social comparison message by increasing their use (e.g., Schultz et al.), we find no evidence that low users or any other subgroups increase their water use, on average, in response to the social comparison message. With regard to mechanisms, the evidence suggests that norm-based messages induce behavioral (variable-cost), rather than technological (fixed-cost), changes in outdoor water use and work through social preferences, rather than by serving as signals of privately efficient behavior to boundedly rational agents. Further, we provide cost-effectiveness evidence of targeting information campaigns.

The next section reviews the most relevant literature. Section 3 describes our methodology. Section 4 describes the experimental design and data. Section 5 provides results and main findings.

2. Experiments with Information-based Environmental Programs

A small number of experimental studies estimate the effects of pro-social and social comparison messages on environmental outcomes. Ayres, Raseman and Shih (2009), Allcott (forthcoming), Costa and Kahn (2010) and Yoeli (2009) focus on energy consumption, while Ferraro and Price and Ferraro, Miranda and Price focus on water consumption. Although the studies differ in terms of location and the content and framing of the messages, the authors find that pro-social messages with social comparisons cause reductions in consumption.[2]

The studies by Ferraro and Price and Ferraro, Miranda and Price are described in section 4. Allcott evaluates a field experiment in Minnesota run by OPower, a firm that promotes energy efficiency for its utility partners. OPower sends home energy reports that include information on strategies to conserve energy, social comparisons of the household's consumption to consumption of geographical neighbors in homes of comparable size, and positive and negative emoticons to indicate the social desirability of the household's position in the distribution used to

[2] Studies on social comparisons without pro-social messages, however, have found no effect (see, for example, the review by Fischer (2008)), suggesting that mixing both types of messages may be necessary.

make the social comparison.[3] OPower's restriction of the social comparison to neighbors with comparable house sizes is intended to heighten the relevance of the social comparison, but it might also reduce the scope and impact of the comparison because consumption variability may be low within the comparison group. Allcott finds that these reports reduce energy consumption by a little over 2 percent. Ayres, Raseman and Shih evaluate two other OPower field experiments (in California and Washington) and find the reports reduce natural gas and electricity use by 1.2 and 2.1 percent, respectively. Yoeli examines a field experiment in California in which the decision to sign-up for a blackout prevention program was randomly varied to be private versus observable to one's neighbors. Participation is 3.6% higher in the publicly observable treatment, but the effect of the program on energy consumption is not reported.

Only three studies examine heterogeneous treatment responses. Allcott runs a quantile regression and Ferraro and Price examine how treatment responses vary as function of being below or above the median use. Both studies report that historically larger users appear to respond more, on average, to social comparison messages. Costa and Kahn use data from the California OPower experiment to test whether responses vary with political affiliation (obtained for a subsample from public voting records). Democratic households, on average, reduce their consumption, whereas Republican households, on average, do not. These three studies do not conduct a careful analysis of heterogeneity and do not examine heterogeneity further. Moreover, these studies do not probe the potential mechanisms through which treatment effects are generated.

3. Methodology

When exploring treatment effect heterogeneity, there is a substantial risk of finding statistically significant differences among subgroups when no true treatment effect heterogeneity exists because the subgroups are formed after the experiment is implemented (Imai and Strauss 2011). To mitigate this potential bias, we adopt five complementary methods to estimate heterogeneous treatment effects. First, prior to the analysis, we select only a few subgroups based on theory, field experience and policy relevance. Second, we demonstrate that, although randomization was

[3] Specifically, treated households receive home energy reports containing: (i) personal use history; (ii) current period neighbor comparison; (iii) twelve-month neighbor comparison; and (iv) energy efficiency advice. For further details: http://www.opower.com/Approach/TargetedMessaging.aspx

not conducted within these subgroups, our large sample size, combined with randomization within 390 small neighborhood strata, generated within-subgroup balance in pre-treatment water use among treated and control households. This balance suggests no systematic bias when drawing inferences from treatment-control outcome contrasts within subgroups. Third, we begin testing for heterogeneity with a nonparametric approach developed by Crump et al. (2008), which tests for the presence of heterogeneity without attempting to characterize the nature of the heterogeneity. Then we impose additional assumptions and estimate quantile treatment effects (Firpo 2007; Bitler et al. 2005, 2006, 2008; Djebbari and Smith; Heckman, Smith and Clements; Heckman, Hienrich and Smith, 2002). Finally, we isolate systematic variations among subgroups through interactions terms between the treatment variable and other covariates. We then use these estimates of heterogeneous subgroup responses, along with complementary analyses, to test hypotheses about the mechanisms through which treatment effects were achieved.

3.1. Heterogeneity in Treatment Responses

To estimate heterogeneous responses, we select covariates that are observable to policymakers and that theory or empirical studies suggest could be important modifiers of the treatment effects. We wish to keep the number of covariates small to avoid charges of data mining. These covariates generate policy-relevant subgroups. We select (i) two measures of previous water use, (ii) three household characteristics, and (iii) two neighborhood characteristics. Furthermore, in our final analysis that uses multiple hypothesis tests, we adjust the Type I error rate for sequential tests. Below, we present the covariates that define subgroups in the order we believe reflects their policy-relevance, and thus the order in which we will conduct the sequential tests. These variables represent our independent variables.

Ferraro and Price's analysis shows that previous water use predicts future water use and provides suggestive evidence of heterogeneous treatment responses conditional on a household's percentile in summer 2006. Moreover, for utilities, previous water use is the easiest characteristic on which to target future messages. We use the two variables used by Ferraro and Price in their treatment effect regressions: June – November 2006 billed use (corresponds to May – October use, which is the main water use season) and April – May 2007 billed use (to reflect changes in landscaping prior to treatment assignment in May 2007).

Mansur and Olmstead (2011) find that, as theory would predict, high-income households in urban areas are less price-sensitive to changes in residential water prices. Whether such households are more or less responsive to *non-pecuniary* approaches is an open empirical question. We cannot observe household income, but we can observe the fair market value of the home in the year in which the treatment was assigned. Based on the high correlation between housing value and income, we use fair market value as a proxy for income (and wealth).[4]

Davis (2010) shows that renters are significantly less likely to have energy efficient appliances, like clothes washers and dishwashers. These results are consistent with the hypothesis that when tenants pay the utility bills, landlords may buy cheap inefficient appliances. In our sample, almost all renters are directly billed (multi-dwelling structures, like apartment buildings, are not in our sample). With regard to water conservation, owner-occupants have a greater incentive to invest in cost-saving water conservation innovations that are capitalized into the value of the home. Owner-occupants may also have greater social connections to their neighbors and thus be more responsive to pro-social messages. Rohe et al. (2001) posit that homeowners are more likely to participate in community activities and might be more civically active because they have higher location-based investments (homes) than renters, higher transaction costs associated with moving, and stronger expectations of staying in their homes longer. DiPasquale and Glaeser (1999) offer evidence that homeowners are more likely to have greater social capital than renters (e.g., homeowners are more likely to participate in solving local problems, and more likely to be members of non-professional organizations).

Owner-occupants, however, may have weaker incentives than renters to reduce water consumption. For example, DiPascuale and Glaeser (1999) also found that homeowners are 12% more likely to garden than non-homeowners. Furthermore, landscaping may be detrimentally affected by water conservation, which could affect home values (landlords may be unable to shift this risk to tenants). Ownership status is revealed by the owner's homestead exemption status (only owner-occupiers receive a homestead exemption).

Another measure that reflects the scope and incentives for water conservation is the age of the home. Older homes, on average, have older water-intensive capital (e.g., toilets), which

[4] The 2007 American Housing Survey shows a high correlation (>0.90) between housing values and incomes, and thus we believe it is reasonable to assume a similar or higher correlation between housing values and wealth.

are more cost-effective to replace to achieve water conservation goals, and they are more likely to have repairable leaks.[5]

Environmental preferences of household occupants would likely also affect their treatment responses. We cannot observe environmental preferences, but survey evidence suggests that environmental preferences vary with education levels and race (e.g. Greenberg, 2005). We (and water utilities) cannot observe education and race at the household-level, and so we use measures of education (percent with bachelor's degree or higher) and race (percent white) at the census block group. The average number of households per block group in our sample is about 425 households.

There are, of course, other covariates that may moderate treatment effects in our experiment, but which we do not measure (e.g., risk preferences). While it may be relevant for theory, this kind of heterogeneity is less relevant for policy makers because it is not easily observed. Recall, we are not making claims that the observable characteristics themselves cause the observed differences in treatment responses. Instead, we wish to measure heterogeneous treatment responses conditional on observable characteristics, with which policymakers can improve program targeting, gain insights into the external validity of experimental results, and better understand the mechanisms through which the treatments operate.

3.2. Nonparametric Tests

Crump et al. (2008) propose tests of two null hypotheses: the conditional average treatment effect is equal to zero (Zero CATE) and the conditional average treatment effect is constant (Constant CATE). Both tests are evaluated using all the subgroup covariates described in 3.1. The Zero CATE null hypothesis states that the impact of a program is zero on average for all subgroups. Testing this hypothesis is relevant for treatment 1 (pure information), which Ferraro and Price report did not have a mean treatment effect different from zero, but which may have had an impact for some subgroups. For each of the other two treatments, which generated nonzero mean impacts, the natural question is whether the treatment effects are constant across subgroups. This question can be evaluated with the Constant CATE test. Crump et al. prove that

[5] We do not explore heterogeneity conditional on lot size because lot size is highly correlated with fair market value.

both tests using nonparametric regression functions based on series estimators can be implemented through regression analysis using an ordinary least squares (OLS) estimator.

The null hypothesis for the Zero CATE test is that the average effect for the subpopulation with covariate values *X* is equal to zero for all values of *X*, while the alternative hypothesis is that the average effect for the subpopulation with covariate values *X* is different from zero for some values of *X*. To test this hypothesis, we run an OLS regression for treated and control group separately controlling for *X*. After obtaining the quadratic form of the difference of estimated coefficients vector ($e(\beta_1 - \beta_0)$), we divide it by the variance-covariance matrix ($e(V_1 + V_0)$). This test statistic follows a chi-square distribution with *k* degrees of freedom:

$$e(\beta_1 - \beta_0)[e(V_1 + V_0)]^{-1} e(\beta_1 - \beta_0) \sim \chi^2(k)$$

The null hypothesis for the Constant CATE test is that the average treatment effect (ATE) for the subpopulation with covariates values *X* is equal to the ATE for all values of *X*, while the alternative hypothesis is that the average effect for subpopulation with covariates value *X* is different from the ATE for some values of *X*. To test this hypothesis, we run an OLS regression with treated and control groups controlling for *X*. After obtaining the quadratic form of the estimated coefficients vector excluding the constant term ($e_{k-1}(\beta_1 - \beta_0)$), we divide it by the variance-covariance matrix excluding the constant term ($e_{k-1}(V_1 + V_0)$). The constant term is excluded because it represents the average effect for everybody. This test statistic follows a chi-square distribution with *k-1* degrees of freedom (because the constant term is excluded):

$$e_{k-1}(\beta_1 - \beta_0)[e_{k-1}(V_1 + V_0)]^{-1} e_{k-1}(\beta_1 - \beta_0) \sim \chi^2(k-1)$$

Following Crump et al. (p.397), we select the final model specification in three ways: (i) include all covariates; (ii) ‘top down’ selection of covariates, where one starts with the full set of covariates and sequentially (one by one) drops the covariate with the smallest t-statistic until all remaining covariates have a t-statistic larger than or equal to 2 in absolute value; and (iii) ‘bottom up’ selection of covariates, where, for each covariate, one runs K regressions with just an intercept and the covariate (K = number of covariates), and then selects from this set the covariate with the highest t-statistic, after which one runs, for each of the remaining covariates, K-1 similar regressions, choosing the one with the highest t-statistic, and continuing the process until no potential covariate has a t-statistic equal to or above 2 in absolute value. We present test results using specifications with higher degree order terms of continuous variables to improve robustness (Crump et al.).[6]

3.3. Quantile Treatment Effects

The nonparametric tests described in 3.2 provide evidence of whether heterogeneous treatment effects exist, but do not characterize this heterogeneity. The next step is to impose some parametric assumptions to begin characterizing this heterogeneity. Within a quantile regression framework, we use estimates of the effects of the treatment on the outcome distribution, or quantile treatment effects (QTE), to infer the presence of heterogeneous treatment effects. QTE show the difference of two marginal distributions at different quantiles, τ_q,

$$\tau_q = F_{Y(1)}^{-1}(q) - F_{Y(0)}^{-1}(q) \qquad (1)$$

rather than the quantile of treatment effect, $\tilde{\tau}_q$,

$$\tilde{\tau}_q = F_{Y(1)-Y(0)}^{-1}(q). \qquad (2)$$

In other words, quantile regressions tell us about the effects on the outcome distribution, rather than on households, which is sufficient to inform us about the presence of heterogeneous treatment effects across quantiles of water use.

Although we do not need an estimate of the effects on households – i.e., the distribution of treatment effects – to infer heterogeneity of causal effects, an estimate of this distribution could be useful to policy makers. For example, one could use the distribution of treatment effects to infer the fraction of the sample for which the treatments increased water use. To estimate this distribution, one needs the joint conditional distribution of treated and untreated states (Heckman, Smith and Clements; Djebbari and Smith 2008). Randomized experiments, however, only provide the marginal distribution of treated outcomes and the marginal distribution of untreated outcomes (which permit the estimation of the ATE).

Nevertheless, under a rank preservation assumption, QTE estimate the distribution of treatment effects (Firpo 2007; Bitler et al. 2005, 2006, 2008). Rank preservation implies that household's ranks in the outcome distribution are the same regardless of whether they are assigned to treatment or control groups (Bitler et al. 2008). If household ranks do not change

[6] The tests were conducted using the command *test_condate* for STATA, which is available at Oscar Mitnik's website: http://moya.bus.miami.edu/~omitnik/

under exposure to the treatment, the ranks in the two marginal distributions from the experiment correspond. Thus, for example, the median outcome in the treated distribution has as its counterfactual the median outcome in the untreated distribution (and so on for all quantiles). In this case, the impact of the treatment on the distribution would be equivalent to the distribution of treatment effects.

Paraphrasing Angrist and Pischke (2009), if we discover, for example, that a message lowers the bottom decile of the water use distribution, we would not necessarily know if someone who would have been a low user without the message is now using less water. We know only that those who use less water with the message are using less water than bottom-decile users would have used without the message. They may not be the same users. In contrast, if the rank preservation assumption holds, the same discovery would mean that the message reduces use among users at the quantile being examined (e.g., the message reduced use among users in the bottom decile).

To test the rank preservation assumption, we follow Bitler et al. (2005) and Djebbari and Smith by using observable covariates of treated and control households. If these covariates vary significantly between treated and control groups in a given quantile, the variation provides evidence against the rank preservation assumption. Dividing our sample into quartiles, we find that 25% of the 84 possible combinations (7 covariates, 4 quartiles and 3 treatments) show statistically significant differences (without adjusting for the multiple sequential tests, which would reduce the number of null hypotheses rejected).[7] These results suggest that some rank reversal may be present based on the covariates selected. In particular, households with high fair market value and high previous water consumption may have migrated down the outcome distribution when treated (See Appendix 2). We conclude there is support for viewing our QTE results as a useful approximation to the distribution of treatment effects, but only as an approximation.

3.4. Subgroup Analysis

The nonparametric and quantile regression approaches described above do not specify which subgroups are most responsive to the treatments. We explore subgroup variation through

[7] Using the same test and data from PROGRESA, the Mexican conditional cash transfer program, Djebbari and Smith (2008) rejected 30% and Lehmann (2010) rejected 31% of the possible combinations.

interactions terms between the treatment variables and other covariates in a regression framework (Heckman, Smith and Clements; Heckman, Heinrich and Smith; Djebbari and Smith).

To guard against spurious findings, we first conduct an F-test to test the null hypothesis that overall that there are no subgroup differences (Type I error rate = 0.05). Then we look at each subgroup in turn, after adjusting the Type I error rate for sequential hypothesis testing with a conservative Bonferroni adjustment (i.e., we take our pre-determined Type I error rate 0.05 and divide it by the number of tests; the null of no difference will only be rejected if $p<0.0075$). As noted earlier, we are estimating causal effects conditional on observable characteristics. We are not making claims that the observable characteristics themselves cause the observed differences in treatment responses.

4. Experimental Design and Data

The Cobb County Water System (CCWS) experiment comprised three treatment groups and one control group:

- Pure Information (Treatment 1): A 'tip sheet' listing different ways to most effectively reduce water use.
- Weak Social Norm (Treatment 2): The 'tip sheet' and a personally addressed letter from CCWS officials encouraging water conservation.
- Strong Social Norm (Treatment 3): The 'tip sheet', the letter from CCWS officials encouraging water conservation, and a social comparison that compared the household's 2006 summer water use to the median County household us. Summer season is from June to September, which is reflected in July to October monthly bills.

In May 2007, the three treatments and control were randomly assigned (mailed) to all residential customers who lived in their homes from May 2006 to April 2007 and used at least 20,000 gallons during the 2006 summer watering season (about 80% of the population). Each treatment consisted of roughly 11,700 houses and the control group consisted of roughly 71,800 houses. For more details about the treatments and experimental design see Ferraro and Price and Ferraro, Miranda and Price. Ferraro and Price find that pure information (treatment 1) had no significant effect, while the weak social norm message (treatment 2) reduced water use by about

2.5%. In contrast, the strong social norm message reduced water use by almost 5%. Ferraro, Miranda and Price (2011) find that only the strong social norm message significantly affects water use three watering seasons after treatment assignment, albeit with smaller effects.

We merged the experimental data with the 2007 County Tax Assessor Database and the 2000 US Census (census block group) using home addresses as the merger link. Tax Assessor data provide relevant information about fair market value, ownership status and the age of the home. The Census provides data on race and education levels. We matched 97% of the experimental sample to the tax assessor data and 89% to the census data.

Table 1 presents descriptive statistics by treatment and control group for the pre-treatment data. Columns (1)-(3) display mean values for households assigned to treatment 1 (pure information), treatment 2 (weak social norm), and treatment 3 (strong social norm). Column (4) displays means for the control group. Column (5) shows the F-statistic and column (6) its respective p-value from a test of the null hypothesis that mean values are equal across treatment and control groups. With the exception of ownership status, for which the mean differences are less than one percentage point, there are no statistically significant differences in pre-treatment variables, including previous water use (recall our sample is over 100,000 observations). These results support the claim that randomization was effective.

The treatments were not, however, randomized within the subgroups identified in 3.1. Nevertheless, given that our sample size is large, our randomization was done within small neighborhood groups and our subgroup set is small, we would expect that observable and unobservable characteristics that affect water use would be well balanced between treatment and control groups within subgroups. To provide evidence of this balance, we examine pre-treatment water use across the treatment and control groups within each subgroup (see Appendix 3). For example, we test (F-test) whether pre-treatment mean water uses across treatment and control groups are statistically indistinguishable from each other within the group of renter-occupied households, then within the group of owner-occupied households, then within the group of above-median fair market value households, then within the group of below-median fair market value households, etc. With sixteen sequential tests and Type I error rate set to 0.05, we would expect approximately one of them to reject the null hypothesis of no difference through chance alone at the $p<0.05$ level. In no test is the null hypothesis rejected.

Table 2 presents descriptive statistics by treatment and control group for post-treatment water use. This analysis replicates and complements the results reported in Ferraro and Price and Ferraro, Miranda and Price with our slightly smaller sample size. Columns (1)-(4) display mean values for treatment and control groups. Columns (5)-(7) display differences with respect to the control group and the statistical significance of these differences. Households consumed less water than the control group in summer 2007, with the difference statistically different from zero for treatment 2 (weak social norm) and treatment 3 (strong social norm).[8] In summer 2008, that difference remains significant only for treatment 3. In summer 2009, none of the treatment effects is significantly different from zero.[9] In winter months, when most water use is indoor water use, only the effect of treatment 3 in 2007/2008 is statistically significant.

5. Results

5.1. Nonparametric Tests

Given treatment 1 had no effect in any year and treatment 2 had no effect beyond 2007, we test whether the effects of these treatments are zero for all subgroups in the relevant years (Zero CATE Test). We also test whether the effect of treatment 2 (for summer 2007) and treatment 3 (for all summers) is constant for all subgroups (Constant CATE Test). For completeness, we report test results for all treatments in all summers. Table 3 summarizes the results for summer water seasons 2007-2009 using the three methods of covariate choice.

Result 1: *There is little evidence of some subgroups responding in all years to treatment 1 (pure information).*

Result 2: *There is weak evidence of heterogeneous responses in 2007 to treatment 2 (weak social norm) and no evidence that any subgroups respond in later years.*

Result 3: *There is strong evidence of heterogeneous responses in all years to treatment 3 (strong social norms).*

[8] Ferraro and Price also show the differences among treatments are statistically different, as is the trend when they are ordered in terms of water use as predicted by their theory (T3<T2<T1<Control).

[9] Ferraro, Miranda and Price increase the statistical precision of these 2008-2009 estimates by estimating a regression model that includes controls for other covariates that contribute to the variability of water use and the randomization strata, and find an effect for treatment 3 in 2009 at the 5% level.

We consider these results in more detail. Each panel shows the results for a specific treatment and each column shows the results for each year. These results represent the number of times that the null hypothesis cannot be rejected (either Zero CATE is equal to zero or Constant CATE is equal to a constant number). For Treatment 1 we evaluate Zero CATE for all years (2007, 2008, 2009). For Treatment 2 we evaluate Constant CATE (2007) and Zero CATE (2008, 2009). For Treatment 3, we evaluate Constant CATE for all years (2007, 2008, 2009).

For treatment 1, results are strong when using higher order covariates. The null hypothesis of zero CATE for all subgroups is rejected ($p<0.05$) all times. These results show evidence that some subgroups might have an average effect different than zero, corroborating Ferraro and Price (forthcoming) results: when excluding top and bottom 0.25 percentile of the distribution, treatment 1 has a statistically significant reduction in water use. However, when running other specification tests (e.g., without higher-order terms and with flexible coding of the covariates as dummy variables) results are very fragile (results not shown). We thus conclude that there is suggestive, but little evidence that treatment 1 affects some subgroups of the experimental sample.

For treatment 2, the null hypothesis of Zero CATE cannot be rejected in 2008 for all three ways of covariate selection suggesting that there is no evidence that some subgroups had an effect different than zero. However the Zero CATE for 2009 is rejected for all tests (but this result is weakened when using dummy variables as covariate specifications where the Zero CATE null hypothesis cannot be rejected –results not shown–). Further, in year 2007 the null hypothesis of Constant CATE is only rejected one time out of six tests. Thus we conclude there is weak evidence of heterogeneous treatment effects for treatment 2.

Regarding treatment 3, the null hypothesis of Constant CATE is rejected all times. This evidence is stronger in year 2007 where all tests are rejected using other covariate specifications (e.g., without higher-order terms and dummy variables).[10] Thus we conclude there is strong evidence of heterogeneous treatments effects for treatment 3.

[10] We also ran all the tests using education and race measured at the census tract. In these tests, all specifications reject the null hypothesis of constant effects for treatment 3 and our results for the other treatments do not change (results not shown).

In summary, using Crump et al.'s (2008) nonparametric tests, we find clear evidence of heterogeneous treatment responses for treatment 3, particularly for year 2007. For treatment 1 and treatment 2, the evidence for heterogeneity is weak and no firm conclusions can be drawn.

5.2. Quantile Regressions

The conclusions drawn from the nonparametric tests combined with quantile regressions depicted in Figures 1-3 provide stronger evidence that the treatment 3 has strong evidence of heterogeneous responses. Figure 1-3 represents the quantile graph for each treatment over the three summer periods. Each graph plots the average treatment effect (dashed line), the QTE (solid line), and the respective confidence intervals of these point estimates.

Result 4: *There is strong evidence of heterogeneous responses only for treatment 3 (strong social norms): water users at the upper end of the distribution respond more.*

For treatment 1 (Figure 1), most of the distribution lies near the zero effect line for all three years without substantial heterogeneity. For treatment 2 (Figure 2), an effect on water use is only detected in 2007, and heterogeneity in this year is confined to the upper half of the distribution. For treatment 3 (Figure 3), there is clear evidence of substantial heterogeneity in 2007, with the greatest water reductions in the upper 20% of the distribution. Summer 2008 also shows greater reduction by high water users, but not as much in previous year. The impacts in 2009 are more homogenous. Thus the results of the quantile regressions are consistent with the nonparametric tests: strong evidence of heterogeneity in responses to treatment 3, and weak or no evidence of such heterogeneity for the other two treatments.

Furthermore, if one were willing to assume rank preservation (see section 3.3) and interpret the figures as estimates of the distribution of treatment effects, one would infer that the treatment messages either reduces water use or has no effect. Nowhere in the distribution is there evidence of statistically significant increases in water use as a result of receiving any treatment message.

5.3. Subgroup Analysis

We define subgroups using the median: a household is thus labeled either as “high value” or “low value.” These subgroups represent policy-relevant subgroups and they are easily identified for policy makers. For ownership status, the subgroups are owners and renters. In Section 4, we demonstrated that pre-treatment mean water use across treatment and control groups are statistically indistinguishable from each other within each subgroup. This result provides evidence that randomization was effective at balancing household characteristics that affect post-treatment water use across the treatment and control groups, permitting the analysis of subgroup treatment effect heterogeneity.

Given the strong evidence in the previous two sections of heterogeneous treatment effects for treatment 3 (strong social norm), and the weak evidence for heterogeneous treatment effects for treatments 1 and 2, we focus on treatment 3. For completeness, we present subgroup analyses for all the treatments, but would caution the reader against interpreting any statistically significant results as evidence of heterogeneous treatment effects for treatment 1 and 2.

In regression models that include the treatments, the subgroups, and all interactions of the treatments and the subgroups (see Appendix 4), we can reject the null hypothesis that the effect of treatment 3 is the same in all subgroups for 2007 ($p<0.001$). For 2008 and 2009, we cannot reject this null hypothesis ($p\approx0.15$). For treatments 1 and 2, we cannot reject this null hypothesis for any year.

Following Heckman, Heinrich and Smith, we simplify the presentation by running independent regressions for each subgroup covariate. Table 4 presents the results for 2007. In the lower panel are the p-values for a hypothesis test of no difference across subgroups within a treatment, unadjusted for repeated hypothesis testing. If we adjust each p-value using the very conservative multiple-hypothesis testing adjustment described in 3.4, we draw the same inferences. We also draw the same inferences from the full regression with interaction terms (see Appendix 4).

Result 5: *For treatment 3 (strong social norms), greater responses are observed in households that use more water in the past, live in more expensive homes and are occupied by owners rather than renters.*

The evidence of heterogeneity across these characteristics is strongest in 2007. In 2008, we observe statistically significant heterogeneity conditional on previous water use and fair market value. In 2009, all the tests fail to reject the null hypothesis of no differences across subgroups (there is some weak evidence of heterogeneity based on fair market value). See Appendix 4-6.

5.4. Targeting Information Campaigns

Ferraro and Price (forthcoming) show that treatment 3 is the most cost-effective treatment among the three treatments tested. They then demonstrate that by targeting only those households at or above the median use for the previous summer, CCWS could obtain 88% of the original reduction for 65% of the original cost.[11] Under this targeting rule, 2007 summer water use would have been expected to decline by approximately 163 million gallons – the equivalent of shutting off the water to about 4500 households — at a cost of $0.43 per thousand gallons reduced.

Could the information from sections 5.1-5.3 be used to further improve targeting? For example, if instead of targeting households based on their use during the previous year's summer, the utility were to instead target households based on their use in the two months before the campaign, it could obtain 80% of the reduction for 48% of the original cost (i.e., a reduction of 149 million gallons at a cost of $0.35 per thousand gallons reduced[12]).

Result 6: *By targeting on households identified as being more responsive to treatment, the water utility can reduce the overall program cost by over 50%, and the cost per gallon reduced by almost 40%, with less than a 20% decline in the total number of gallons reduced.*

If the utility were willing to sacrifice further reductions in use to achieve greater cost-effectiveness, combinations of targeting could further increase cost-effectiveness towards $0.30 per thousand gallons reduced (e.g., target large water users who own their home and who reduce, on average, over 3,000 gallons/household). Combined with information on the benefits from reduced water use, heterogeneous treatment effect estimates could be used to determine an optimal targeting strategy.

[11] Recall that approximately the bottom quintile of water consumers is not part of the experiment.

[12] When contrasting cost-effectiveness across different conservation and augmentation policy options, one must remember to also consider the persistence of the treatment effects (Ferraro, Miranda and Price, 2011).

5.5. Mechanisms of the Strong Social Norm Message (Treatment 3)

Among the experimental treatments, treatment 3 had the largest and most persistent effect, and the strongest evidence of heterogeneous treatment effects. In this section, we explore the mechanisms through which this treatment affects household behavior. We present evidence with regard to three mechanism hypotheses:

(i) The treatment effects are driven mainly by continuous, behavioral changes with recurring costs (e.g., watering outdoors less frequently or washing full loads of laundry or dishes) rather than one-time, behavioral or technological investments (e.g., fixing leaks, buying new appliances)

(ii) The treatment effects are driven mainly by changes in outdoor water use rather than changes in indoor water use; and,

(iii) The social comparison in the message affects behavior by highlighting social norms rather than by sending signals about privately efficient behavior (i.e., highlighting cost-savings opportunities).

The first hypothesis is relevant for understanding the long-term effects of the program. This hypothesis can also be viewed as asking whether the home is treated or the home dwellers are treated. If the home were treated (e.g., a leak fixed; an efficient irrigation system installed), one would expect on-site treatment effects to persist, even after the current inhabitants depart.

The second hypothesis is relevant to understanding the environmental effects of the treatment. Most of the indoor water used in Cobb County returns to the surface water system from which it was drawn. Because of processes like evapotranspiration and infiltration, most of the outdoor water used does not return on a time scale relevant for stream flow. Previous empirical work implies outdoor water use is more price elastic (e.g., Mansur and Olmstead 2011). Thus, one might predict that, after receiving a message, households would first look to reduce water from outdoor use, just as they would respond to a price increase.

The third hypothesis has not, to our knowledge, been raised in the literature on social comparisons. Rather than working through social preferences, as assumed in the literature, the social comparison may work simply by conveying costly information about private costs and benefits. In an incomplete-information world with costly information acquisition or boundedly rational agents, households may not be optimizing their water use. The lack of a treatment effect from the information-only treatment (treatment 1) suggests that households already know how they can reduce water use. They may assume, however, that adopting (or disadopting) these practices would not be utility-maximizing given their beliefs about costs and benefits. Yet when confronted with information about others' water use, they may update their beliefs (e.g., "I didn't know there could be gains from adopting these tips until I saw how my use compared to others' use."). Thus, the "social" comparison may actually be a "private" signal. Rather than harnessing pro-social preferences, the comparison helps self-interested, utility-maximizing agents get closer to the privately optimal water use pattern under complete information. We emphasize the social comparison, rather than treatment 3 in its entirety, because (1) Ferraro and Price showed that the tip sheet had no detectable effect and that there were statistically significant differences across treatments, and (2) treatment 2 could only have affected behavior through social preferences. Thus we can conclude that *some* of the effect from treatment 3 arose from social preferences. The question that remains is "How did the addition of the social comparison reduce water use further?"

5.5.1. Recurring behavioral changes versus one-shot technological investments

To examine the first hypothesis about the nature of the actions taken by households, we use three pieces of information. First, if households were to reduce water use mainly through one-shot investments (e.g., fix leaks, install low-flow toilets), one would expect relatively constant treatment effects across years within seasons. Yet, as indicated in Table 5, the effects wane over time within season (the same waning occurs also in spring months). However, inter-year variations in other factors could also explain this pattern.

Second, if households were to reduce water use mainly through one-shot, fixed-cost investments, one would expect such investments to be more likely in older houses where such investments are more cost-effective (e.g., they are more likely to have leaking pipes and older appliances). Table 4, however, shows that there is no difference in the responses between older and newer houses.

A final test exploits the re-framing of the first hypothesis in terms of asking whether the home or home dweller is treated. If one-shot, fixed-cost investments were driving reductions, the treatment effects should not disappear when the message recipients move out of their homes. We define “movers” as those households where the customer identification number changed between December 2007 and September 2008. Table 6 shows that, in summer 2007, movers (who had not yet moved) and non-movers reacted similarly to treatment 3. In fact movers reduced a bit more than non-movers: 1,900 versus 1,700 gallons (the difference is not statistically different from zero in a pooled regression model). In summer 2008, however, treatment 3 had a *positive*, but statistically insignificant, effect on households in which the message recipients had moved out (the difference between movers and nonmovers in a pooled regression model is statistically significant at p=0.06). Together, these three pieces of evidence suggest a seventh result:

Result 7: *The evidence is consistent with the hypothesis that the effects of the strong social norm message (treatment 3) are driven mainly by behavioral changes with recurring costs rather than fixed-cost investments in technology.*

5.5.2. Outdoor versus indoor water use changes

To examine the second hypothesis, one would ideally be able to observe outdoor and indoor water use separately, but Cobb County does not measure these uses separately. So we must depend on theory and previous empirical evidence suggesting that outdoor water demand is more price elastic, and indirect empirical evidence from our experiment that shows treatment effects are largest in the months in which outdoor watering is typically observed.

The lowest consumption in Cobb County occurs in December, during the winter when, water utility employees say, most use is indoor use. The highest consumption occurs in July, during the summer when most outdoor watering occurs. In 2006, before the experiment was implemented, the average December consumption in our sample was 6007 gallons and the average July consumption was 11,470, a 90% increase. We thus believe a contrast of treatment effects in December and July captures differences in indoor versus outdoor use.

Table 7 presents estimates of the average treatment effects for July 2007, December 2007, and July 2008 from regressions that include the strata in which the randomization was conducted (meter routes) and pre-treatment water use variables. In July 2007 and July 2008, treatment 3 has large and statistically significant effects on water use, and the effect in July 2008 is half of the effect in July 2007. In December 2007, however, the effect of treatment 3 is small and statistically insignificant. Comparing coefficients across regressions, the treatment effects in

July 2007 or July 2008 versus December 2007, the differences are statistically significant ($p<0.01$ and $p<0.04$, respectively). In order to argue that these results do not imply most of the treatment effect was coming from outdoor watering, one would have to argue that waning only occurs with indoor water use (and thus the July 2008 effect is the same as the outdoor effect in July 2007). Such an argument is difficult to maintain for two reasons. First, about 60% of indoor water use is for toilets, washing and bathing,[13] and it is hard to imagine how behavioral changes indoors (rather than technological changes) could have accounted for half the water reduction observed in July 2007. Second, the treatment effect from July 2009 is small and insignificantly different from zero (-0.081, $p=0.23$) and significantly different from the treatment effect in July 2008 ($p=0.06$). Thus one would have to argue that waning in the outdoor treatment effects only started after July 2008. We therefore believe that the evidence supports an eighth result:

Result 8: *The empirical evidence is consistent with the hypothesis that the effects of the strong social norm message (treatment 3) are driven mainly by changes in outdoor water use.*

5.5.3. Social versus private preferences

To examine the hypothesis that the social comparison works through social preferences rather than private preferences, we use three pieces of information. First, we take advantage of Cobb County's increasing block price structure and test for the presence of private preference-based mechanisms. The price per additional gallon of water use increased at two thresholds: 9,000 gallons/month (from \$2.21 to \$2.55/1,000 gallons) and 16,000 gallons/month (to \$2.88/1,000 gallons). If private preferences motivate treatment responses, households that were using just above these threshold limits in the pre-treatment summer period should, on average, be more likely to reduce their water use post-treatment than households just below the threshold limits (because the expected cost savings are higher for households above the threshold).

Using bandwidths of 500, 1000, 2000, 3000 and 4000 gallons around the threshold, we test whether households just above the threshold respond more to the message than households just below the threshold within the bandwidth. We define "above the threshold" in two ways: as a dummy variable and as a continuous variable (the difference between water use in summer

[13] http://www.epa.gov/WaterSense/pubs/indoor.html

2006 and the threshold). We estimate models both with and without the other subgroup-defining covariates. We also re-estimate all the models using only households who were consistently above or below the threshold every month in the previous summer (rather than above or below based on average monthly consumption during the summer). Thus we estimate 52 regression models (see Appendix 7 – 9). In only three of them could we reject the null hypothesis of no difference between those above and below the threshold within the bandwidth, and in these cases the sign of the estimated coefficient was inconsistent with the hypothesis (in fact, for 34 of the 52 cases, the estimated coefficient was inconsistent with the hypothesis).

Next we combine the results from 5.3 and 5.4.1 with assumptions about heterogeneous responses when private preferences motivate water use reductions. In 5.4.1, we presented evidence consistent with the hypothesis that households respond to the treatment with behavioral changes that have recurring costs. If this hypothesis were true *and* water reductions were driven by private preferences, renters and owners should be equally likely to decrease water use to save money because savings are immediate and not capitalized into the value of the house, as they might be with one-shot, technology investments. In contrast, we observe in 5.3 that owners are more responsive to the treatment (even after conditioning on all the other covariates; see Appendix 4).

Third, we assert that if private preferences for cost savings were driving reductions in water use, then after holding owner/renter status constant, a change in the percentage of renters in the neighborhood (census block group) should not influence the response of a household to treatment 3. The political science and economics literature cited in section 3.1, however, suggests this percentage may affect a household for whom the social comparison is working through social preferences because neighborhoods with greater ownership rates have greater social connections and thus the norm created by the social comparison to neighbors is more relevant or salient. Moreover, we would not expect any such interaction between the proportion of households renting in a neighborhood and treatment 1 or treatment 2 because there is no social comparison in these treatments. Thus keeping these treatments in the model helps protect against spurious rejections of the null hypothesis because of other factors that may be correlated with an increase in renters in a neighborhood. We further protect against such rejections by adding previous water use, home characteristics and neighborhood fixed effects (meter route strata) to the model.

We define high-rental neighborhoods as those who have a percentage of renters greater than the median. We then create interaction terms with a dummy variable for above-median proportion of renters and the treatment dummy variables. Results are shown in Table 8. The interaction term is positive and significantly different from zero for only treatment 3. Holding ownership status, home characteristics, previous water use and other neighborhood characteristics constant, households who receive treatment 3 are more responsive in census block groups with low percentages of renters.

Recall that treatment 3 augments treatment 2 with a social comparison treatment, and treatment 2 affects water use significantly less than treatment 3. Combining this observation with the evidence in this subsection, we arrive at our final result:

Result 9*: The evidence is consistent with the hypothesis that the social comparison induces greater water use reductions by highlighting social norms rather than by sending signals about privately efficient behavior.*

6. Conclusions

Non-pecuniary, information-based strategies are increasingly being used to influence individual decision-making to achieve policy objectives. Despite the increasing application of these strategies, their conceptual and empirical foundations remain under-researched. Although a growing number of scholars are conducting randomized experiments to test information-based strategies, many of them only report average treatment effects, thereby ignoring variation in the treatment effects. Moreover, none have attempted to elucidate the mechanisms through which these strategies operate.

In experimental studies in which treatments are not randomized within subgroups, or in any observational study, one must be cautious when estimating heterogeneous treatment effects. Unlike many studies of heterogeneous treatment effects in social experiments, we reduce the risk of mislabeling spurious correlations as heterogeneous treatment effects by combining complementary empirical approaches in an experiment with a large sample size and randomization conducted within small neighborhood strata. These attributes afford us better statistical power and experimental control in estimating heterogeneous treatment effects across observable subgroups in the population.

In our study of an information-based environmental program aimed at inducing voluntary reductions in the use of a common pool resource, we find strong evidence of heterogeneous treatment effects for a message that augments pure information with pro-social language and social comparisons (strong social norm message). In contrast, the evidence of heterogeneous treatment effects from pure information alone or pure information with pro-social language but no social comparison (i.e., traditional conservation messages) is weak. The social psychology literature on social norms predicts heterogeneous responses to social comparison messages (e.g., Schultz et al.), but this predicted heterogeneity is in the form of a "boomerang" effect, whereby low users discover through the social comparison that they are low users and, in response, increase their use. Based on this literature, OPower's norm-based, energy conservation program (section 2) supplements its social comparisons with emoticons: if a household's energy use is below the average of its comparison group, it receives a green "smiley face" that is assumed to prevent the boomerang effect.

Despite the absence of emoticons in our experiment, we find no evidence of a boomerang effect. Assuming rank preservation is a good approximation in our study (section 3.3), there is no evidence of statistically significant increases in water use anywhere in the distribution as a result of receiving the social comparison message. Likewise, our subgroup analysis reveals no evidence that any subgroup, on average, increases its water consumption as a result of receiving the message. Complementary evidence comes from Ferraro and Price who show that, on average, below-median users reduce their water use upon receiving the social comparison, rather than increase it.

Turning to our mechanism hypotheses, the evidence suggests that the strong social norm message operates through behavioral changes with recurring variable costs rather than one-shot, fixed-cost investments, and through changes in outdoor watering rather than indoor watering. We also explore a third mechanism hypothesis that posits a rival explanation of how social comparisons affect behavior: rather than operating through social preferences, they may simply convey costly information about privately efficient behavior to households with incomplete information. The evidence, however, is inconsistent with this rival explanation. Social comparisons do seem to work through social preferences. A better understanding of the mechanisms through which norm-based messages operate is important for future attempts to estimate the full welfare implications of information-based policies and programs.

Finally, our study has at least three policy implications. The first relates to the external validity of the CCWS experiment: sites with poor households, many renters, or low water use may not see as large of an impact as Cobb County did from an information campaign that augments information and pro-social language with social comparisons.

Second, making information about heterogeneous treatment effects available to decision makers can greatly improve program cost-effectiveness. We demonstrated that with improved targeting based on observable household characteristics, the overall costs of the program could be reduced by over 50% with less than a 20% decline in the aggregate impact.

Third, as suggested by Ferraro and Price, pecuniary and norm-based, non-pecuniary policies may be complementary. The strong social norm message had an immediate effect on water use in the month after the message was sent and high-income households are most responsive to the message. Thus in contrast to water conservation programs that use pecuniary incentives, for which average responses are slow and high-income households are least responsive (e.g. Mansur and Olmstead), programs based on norm-based incentives work quickly and are most effective among high-income households. Moreover, price changes are typically expected to lead to a persistent change in the quantity demanded, whereas the evidence from this experiment suggests the effects of norm-based approaches wane over time. Thus the two approaches may be preferred in different contexts (e.g., a need to change short-term demand rather than long-term demand) and may be complementary when combined. Future experiments should directly test the hypothesis of complementarity between the two approaches by randomly assigning pecuniary and non-pecuniary incentives, in isolation and in combination (and, if possible, randomizing within subgroups of interest).

7. References

Angrist, Joshua and Jörn-Steffen Pischke (2009), Mostly Harmless Econometrics: An Empiricist's Companion, Princeton University Press.

Allcott, Hunt (forthcoming), Social Norms and Energy Conservation, *Journal of Public Economics*.

Allcott, Hunt and Sendhil Mullainathan (2010), Behavioral and Energy Policy, *Science*, 5 March 2010, Vol. 327, No. 5970, pp. 1204–1205.

Angrist, Joshua (2004), Treatment Effect Heterogeneity in Theory and Practice, *Economic Journal*, Vol. 114, March, pp. C52–C83.

Ayres, Ian; Sophie Raseman and Alice Shih (2009), Evidence from two large field experiments that peer comparison feedback can reduce residential energy usage, NBER Working Paper # 15386.

Bennear, Lori and Sheila Olmstead (2008), The impacts of the ''right to know'': Information disclosure and the violation of drinking water standards, *Journal of Environmental Economics and Management*, Vol. 56, pp. 117–130.

Bitler, Marianne; Jonah Gelbach and Hilary Hoynes (2005), Distributional impacts of the Self-Sufficiency Project, NBER Working Paper # 1626.

Bitler, Marianne; Jonah Gelbach and Hilary Hoynes (2006), What Mean Impacts Miss: Distributional Effects of Welfare Reform Experiments, *American Economic Review*, Vol. 96, No. 4, pp. 988–1012.

Bitler, Marianne; Jonah Gelbach and Hilary Hoynes (2008), Distributional impacts of the Self-Sufficiency Project, *Journal of Public Economics*, Vol. 92, pp. 748–765.

Bjørner, Thomas B., Lars G. Hansen and Clifford S. Russell (2004), Environmental Labeling and Consumers' Choice: an empirical analysis of the effect of the Nordic Swan, *Journal of Environmental Economics and Management*, Vol. 47, No. 3, pp.411-434.

Bui, Linda and Christopher Mayer (2003), Regulation and Capitalization of Environmental Amenities: Evidence from the Toxic Release Inventory in Massachusetts, *The Review of Economics and Statistics,* Vol. 85, No. 3, pp. 693–708.

Costa, Dora and Matthew Kahn (2010), Energy Conservation 'Nudges' and Environmentalist Ideology: Evidence from a Randomized Residential Electricity Field Experiment, NBER Working Paper # 15939.

Crump, Richard; Joseph Hotz; Guido Imbens and Oscar Mitnik (2008), Nonparametric Tests for Treatment Effect Heterogeneity, *The Review of Economics and Statistics*, August 2008, Vol. 90, No. 3, pp. 389–405.

Davis, Lucas (2010), Evaluating the Slow Adoption of Energy Efficient Investments: Are Renters Less Likely to Have Energy Efficient Appliances?, NBER Working Paper # 16114.

Deaton, Angus (2010), Instruments, Randomization, and Learning about Development, *Journal of Economic Literature*, Vol. 48, No. 2, pp. 424–455.

DiPasquale, Denise and Edward Glaeser (1999), Incentives and Social Capital: Are Homeowners Better Citizens? *Journal of Urban Economics*, Vol. 45, pp. 354–384.

Djebbari, Habiba and Jeffrey Smith (2008), Heterogeneous Impacts in PROGRESA, *Journal of Econometrics*, Vol. 145, pp. 64–80.

Duflo, Esther and Emmanuel Saez (2003), The Role of Information and Social Interactions in Retirement Plan Decisions: Evidence from a Randomized Experiment, *The Quarterly Journal of Economics*, Vol. 118, No. 3, pp. 815–842.

Ferraro, Paul and Michael Price (forthcoming), Using Non-Pecuniary Strategies to Influence Behavior: Evidence from a large-scale field experiment, *The Review of Economics and Statistics*.

Ferraro, Paul, Juan Jose Miranda and Michael Price (2011), The Persistence of Treatment Effects with Non-Pecuniary Policy Instruments: Evidence from a Randomized Environmental Policy Experiment, *American Economic Review: Papers and Proceedings*, Vol. 101, No. 3: 318–322.

Firpo, Sergio (2007), Efficient Semiparametric Estimation of Quantile Treatment Effects, *Econometrica*, Vol. 75, No. 1, pp. 259–276.

Fischer, Corinna (2008), Feedback on Household Electricity Consumption: a Tool for Saving Energy?, *Energy Efficiency*, Vol. 1, No. 1, pp. 79–104.

Goldstein, Noah J., Robert B. Cialdini, and Vladas Griskevicius (2008), A Room with a Viewpoint: Using Social Norms to Motivate Environmental Conservation in Hotels, *Journal of Consumer Research*, 35, pp. 472 – 482.

Greenberg, Michael (2005), Concern about Environmental Pollution: How Much Difference Do Race and Ethnicity Make? A New Jersey Case Study, *Environmental Health Perspectives*, Vol. 113, No. 4, pp. 369–374.

Greenstone, Michael, and Ted Gayer (2009), Quasi-experimental and Experimental Approaches to Environmental Management, *Journal of Environmental Economics and Management* Vol. 57, No. 1, pp.21-44.

Habyarimana, James and William Jack (2011), Heckle and Chide: Results of a randomized road safety intervention in Kenya, Journal of Public Economics, Vol. 95, pp. 1438–1446.

Heckman, James; Jeffrey Smith and Nancy Clements (1997), Making the Most Out of Programme Evaluations and Social Experiments: Accounting for Heterogeneity in Programme Impacts, *Review of Economic Studies*, Vol. 64, pp. 487–535.

Heckman, James, Carolyn Heinrich and Jeffrey Smith (2002), The Performance of Performance Standards, *Journal of Human Resources*, Vol. 37, No. 4, Autumn 2002, pp. 778–811.

Hotz, Joseph, Guido Imbens and Julie Mortimer (2005), Predicting the efficacy of future training programs using past experiences at other locations, *Journal of Econometrics*, Vol. 125, pp. 241–270.

House of Lords, Science and Technology Select Committee (2011), Behaviour Change, 2nd Report of Session 2010-12, London: The Stationary Office Limited.

Imai, Kosuke and Aaron Strauss (2011), Estimation of Heterogeneous Treatment Effects from Randomized Experiments, with Application to the Optimal Planning of the Get-Out-the-Vote Campaign, *Political Analysis*, Vol. 19, No. 1 (Winter), pp. 1-19.

Imbens, Guido and Jeffrey Wooldridge (2009), Recent Developments in the Econometrics of Program Evaluation, *Journal of Economic Literature*, Vol. 47, No. 11, pp. 5–86.

Jin, Ginger and Phillip Leslie (2003), The Effect of Information on Product Quality: Evidence from Restaurant Hygiene Grade Cards, *Quarterly Journal of Economics*, Vol. 118, No. 2, pp. 409–451.

Mansur, Erin and Sheila Olmstead (2011), The Value of Scarce Water: Measuring the Inefficiency of Municipal Regulations, Yale University, Working Paper.

Lehmann, Michael-Christian (2010), Spatial Externalities of Social Programs: Why Do Cash Transfer Programs Affect Ineligible's Consumption?, Paris School of Economics, Working Paper.

List, John A. and Michael K. Price (forthcoming), Using Field Experiments in Environmental and Resource Economics, *Review of Environmental Economics and Policy*

Rohe, William, Shannon Van Zandt and George McCarthy (2001), The Social Benefits and Costs of Homeownership: A Critical Assessment of the Research, Low-Income Homeownership Working Paper Series LIHO-01.12, Joint Center for Housing Studies, Harvard University.

Schultz, P. Wesley, Jessica M. Nolan, Robert B. Cialdini, Noah J. Goldstein, and Vladas Griskevicius. 2007. The Constructive, Destructive, and Reconstructive Power of Social Norms, *Psychological Science*, 18, pp. 429 – 434.

Smith, Kerry and William Desvousges (1990), Risk Communication and the Value of Information: Radon as a Case Study, *The Review of Economics and Statistics*, Vol. 72, No. 1, pp. 137–142.

Smith, Kerry; William Desvousges; Reed Johnson, and Ann Fisher (1990), Can Public Information Programs Affect Risk Perceptions? *Journal of Policy Analysis and Management*, Vol. 9, No. 1, pp. 41–59.

Thaler, Richard and Cass Sustein (2008), Nudge. Improving Decisions About Health, Wealth, and Happiness, Penguin Books.

Yoeli, Erez (2009), Does Social Approval Stimulate Prosocial Behavior? Evidence from a Field Experiment in the Residential Electricity Market, University of Chicago, Working Paper.

Table 1
Pre-Treatment Descriptive Statistics

Variable	(1) Technical Advice (T1)	(2) Weak Social Norm (T2)	(3) Strong Social Norm (T3)	(4) Control	(5) F-Statistic	(6) p-value
Pre-Treatment Data						
Water Consumption Jun-Nov 2006 1/	58.286	58.012	58.381	58.142	0.200	0.897
Water Consumption Apr-May 2007 1/	15.952	15.841	15.957	15.867	0.380	0.768
House's Fair Market Value	257,824	260,984	260,888	258,647	0.950	0.415
Age of House	20.753	20.830	20.710	20.723	0.230	0.878
% Owner Occupiers	0.844	0.836	0.835	0.844	3.190	0.023
% Population 25 years ≥ Bachelor 2/	0.728	0.727	0.728	0.728	0.020	0.997
% of Households White 2/	0.842	0.842	0.843	0.842	0.160	0.924

1/ In thousands of gallons.

2/ At census block group level.

Sources: Experimental Data, 2007 Cobb County Tax Assessor Database, 2000 US Census.

Table 2
Post-Treatment Water Use Descriptive Statistics
(in thousand of gallons)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Variable	Technical Advice (T1)	Weak Social Norm (T2)	Strong Social Norm (T3)	Con-trol	Diff (1)-(4)	Diff (2)-(4)	Diff (3)-(4)
Post-treatment Data							
Summer 2007 1/	36.35	35.39	34.87	36.40	-0.05	-1.00 ***	-1.53 ***
Summer 2008 1/	25.51	25.33	24.99	25.49	0.02	-0.17	-0.50 **
Summer 2009 1/	27.78	27.38	27.18	27.42	0.36	-0.04	-0.24
Winter 07/08 2/	21.63	21.58	21.43	21.71	-0.08	-0.13	-0.28 **
Winter 08/09 2/	21.83	21.57	21.63	21.79	0.04	-0.22	-0.16

1/ Summer season comprises July to October use.

2/ Winter season comprises December to March use.

Source: Experimental Data.

Table 3
Test of Zero CATE & Constant CATE
(Covariates with Higher Order Terms)

	Treatment 1		
	2007	2008	2009
	(Zero CATE)	(Zero CATE)	(Zero CATE)
Top Down Selection of Covariates	2	2	2
Bottom Up Selection of Covariates	2	2	2
All Covariates	2	2	2
% Rejections	*100.0%*	*100.0%*	*100.0%*
	Treatment 2		
	2007	2008	2009
	(Constant CATE)	(Zero CATE)	(Zero CATE)
Top Down Selection of Covariates	1	0	2
Bottom Up Selection of Covariates	0	0	2
All Covariates	0	0	2
% Rejections	*16.7%*	*0.0%*	*100.0%*
	Treatment 3		
	2007	2008	2009
	(Constant CATE)	(Constant CATE)	(Constant CATE)
Top Down Selection of Covariates	2	2	2
Bottom Up Selection of Covariates	2	2	2
All Covariates	2	2	2
% Rejections	*100.0%*	*100.0%*	*100.0%*

Note: These results are summarized from Appendix 1. They represent the number of times that the null hypothesis cannot be rejected. For Treatment 1 we evaluate Zero CATE (2007, 2008, 2009). For Treatment 2 we evaluate Constant CATE (2007) and Zero CATE (2008, 2009). For Treatment 3, we evaluate Constant CATE (2007, 2008, 2009).

Table 4
Subgroup Analysis for Water Consumption (Summer 2007)

	Dependent Variable: Summer 2007						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Previous water use (Jun - Nov 2006)	Previous water use (Apr - May 2007)	Fair Market Value	Owners / Renters 1/	Age of Home	% White	% with higher degree
Treatment 1	-0.320	-0.136	0.171	0.399	0.121	-0.550	-0.551
(Technical Advice)	(0.171)	(0.196)	(0.264)	(0.976)	(0.472)	(0.342)	(0.342)
Treatment 2	-0.444*	-0.559**	-0.732**	-0.310	-0.651	-0.678	-0.925*
(Weak Social Norm)	(0.185)	(0.197)	(0.268)	(0.777)	(0.442)	(0.352)	(0.371)
Treatment 3	-0.653**	-0.722**	-0.772**	0.248	-1.533**	-1.421**	-1.100**
(Strong Social Norm)	(0.159)	(0.198)	(0.251)	(0.754)	(0.405)	(0.322)	(0.340)
Subgroup var.	27.45**	27.22**	16.54**	2.634**	-4.275**	9.694**	8.642**
(high = 1)	(0.200)	(0.208)	(0.212)	(0.341)	(0.221)	(0.227)	(0.228)
Treat1*high subgrop var.	0.445	0.420	-0.415	-0.530	-0.294	0.894	0.876
	(0.559)	(0.589)	(0.595)	(1.027)	(0.616)	(0.628)	(0.630)
Treat2*high subgrop var.	-0.774	-0.614	-0.383	-0.802	-0.773	-0.775	-0.296
	(0.519)	(0.541)	(0.552)	(0.835)	(0.570)	(0.587)	(0.588)
Treat3*high subgrop var.	-1.994**	-2.176**	-1.406**	-2.101**	0.0419	-0.334	-0.887
	(0.485)	(0.503)	(0.523)	(0.808)	(0.543)	(0.555)	(0.560)
Constant	23.02**	23.87**	28.11**	34.17**	38.51**	31.28**	31.82**
	(0.0636)	(0.0746)	(0.0965)	(0.320)	(0.165)	(0.130)	(0.134)
Observations	102,887	102,887	102,871	102,869	102,461	94,833	94,833
R-squared	0.222	0.217	0.080	0.001	0.006	0.029	0.023
p-value equal impact T1	0.426	0.476	0.486	0.606	0.633	0.154	0.164
p-value equal impact T2	0.136	0.257	0.488	0.337	0.175	0.187	0.614
p-value equal impact T3	0.000	0.000	0.007	0.009	0.939	0.547	0.113

Note: All water consumption variables are in thousands of gallons.

1/ In the case of Owners (=1) / Renter (=0), interaction terms are for owner group rather than high group.

Robust standard errors in parentheses

** $p<0.01$, * $p<0.05$

Table 5
Linear Regressions of Water Seasons
(with meter route fixed effects)

	(1) Summer07	(2) Winter0708	(3) Summer08	(4) Winter0809	(5) Summer09
Treatment 1 (Technical Advice)	-0.237	-0.121	-0.0702	0.0335	0.241
	(0.189)	(0.121)	(0.166)	(0.189)	(0.169)
Treatment 2 (Weak Social Norm)	-0.991**	-0.108	-0.190	-0.239	-0.0604
	(0.171)	(0.122)	(0.184)	(0.181)	(0.167)
Treatment 3 (Strong Social Norm)	-1.741**	-0.359**	-0.637**	-0.223	-0.344*
	(0.166)	(0.129)	(0.161)	(0.181)	(0.162)
Water Use from Jun - Nov 2006	0.347**	0.0258**	0.127**	0.0379**	0.170**
	(0.0130)	(0.00485)	(0.00981)	(0.00913)	(0.0106)
Water Use in Apr and May 2007	0.829**	0.335**	0.414**	0.237**	0.427**
	(0.0450)	(0.0171)	(0.0246)	(0.0160)	(0.0251)
Constant	1.874	16.21**	7.086**	15.79**	14.14**
	(1.595)	(0.736)	(0.841)	(0.720)	(1.156)
Observations	106,669	106,669	106,669	106,669	106,669
R-squared	0.634	0.129	0.248	0.021	0.333

Note: All water consumption variables are in thousands of gallons. Winter season runs from December to March billing. Summer season runs from July to October billing.
Robust standard errors in parentheses
** p<0.01, * p<0.05

Table 6
Linear Regressions: Movers and Non-Movers
(with meter route fixed effects)

	(1)	(2)	(3)	(4)
	Summer 2007		Summer 2008	
	Mover 1/	Non Mover	Mover 1/	Non Mover
Treatment 1 (Pure Information)	-1.938	-0.127	-1.367	-0.00830
	(1.058)	(0.192)	(0.968)	(0.169)
Treatment 2 (Weak Social Norm)	-1.329	-0.970**	-0.357	-0.175
	(1.051)	(0.174)	(1.104)	(0.187)
Treatment 3 (Strong Social Norm)	-1.931*	-1.695**	0.826	-0.671**
	(0.959)	(0.169)	(0.999)	(0.163)
Water Use from June - November 2006	0.226**	0.352**	0.124**	0.126**
	(0.0293)	(0.0133)	(0.0322)	(0.0102)
Water Use in April and May 2007	0.985**	0.817**	0.133*	0.423**
	(0.0872)	(0.0462)	(0.0587)	(0.0259)
Constant	-12.94*	2.679	6.177*	7.292**
	(5.905)	(1.622)	(2.424)	(0.861)
Observations	3,667	102,811	3,667	102,811
R-squared	0.525	0.640	0.216	0.254

Note: All water consumption variables are in thousands of gallons.
1/ New residents between December 2007 – September 2008.
Robust standard errors in parentheses
** p<0.01, * p<0.05

Table 7
Linear Regressions of July 2007, December 2007 and July 2008
(with meter route fixed effects)

	(1) July 2007	(2) December 2007	(3) July 2008
Treatment 1 (Technical Advice)	-0.0535	-0.0495	-0.0673
	(0.0618)	(0.0453)	(0.0623)
Treatment 2 (Weak Social Norm)	-0.334**	-0.0572	-0.0542
	(0.0618)	(0.0453)	(0.0623)
Treatment 3 (Strong Social Norm)	-0.548**	-0.0870	-0.220**
	(0.0618)	(0.0453)	(0.0623)
Water Use from June - November 2006	0.103**	0.00763**	0.0360**
	(0.000652)	(0.000478)	(0.000658)
Water Use in April and May 2007	0.228**	0.0927**	0.119**
	(0.00221)	(0.00162)	(0.00223)
Constant	-0.899*	-1.215**	3.481**
	(0.386)	(0.283)	(0.389)
Observations	106,669	106,669	106,669
R-squared	0.555	0.150	0.193

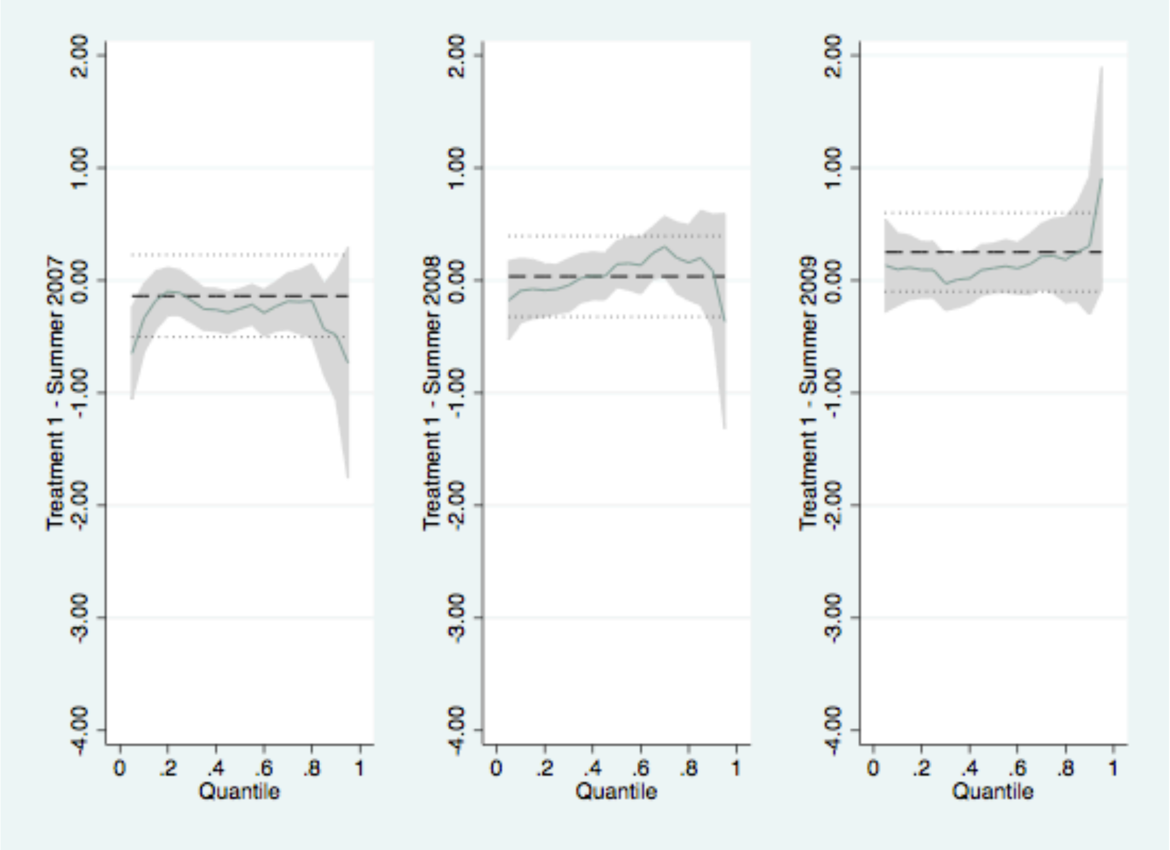
Standard errors in parentheses
** p<0.01, * p<0.05

Table 8
Social Norms or Signal of Privately Optimal Behavior
(with meter route fixed effects)

	Summer 2007
Treatment 1 (Pure Information)	-0.0282
	(0.320)
Treatment 2 (Weak Social Norm)	-0.898**
	(0.265)
Treatment 3 (Strong Social Norm)	-2.127**
	(0.264)
High Proportion Renters (=1 if > median)	0.113
	(0.282)
High Proportion Renters * Treat 1	-0.219
	(0.396)
High Proportion Renters * Treat 2	-0.144
	(0.356)
High Proportion Renters * Treat 3	0.888*
	(0.351)
Water Use from June - November 2006	0.335**
	(0.0144)
Water Use in April and May 2007	0.823**
	(0.0517)
Ownership status	0.701**
	(0.189)
Fair Market Value	1.79e-05**
	(3.12e-06)
Age of Home	0.0265*
	(0.0112)
Constant	-2.422
	(1.799)
Observations	95,233
R-squared	0.638

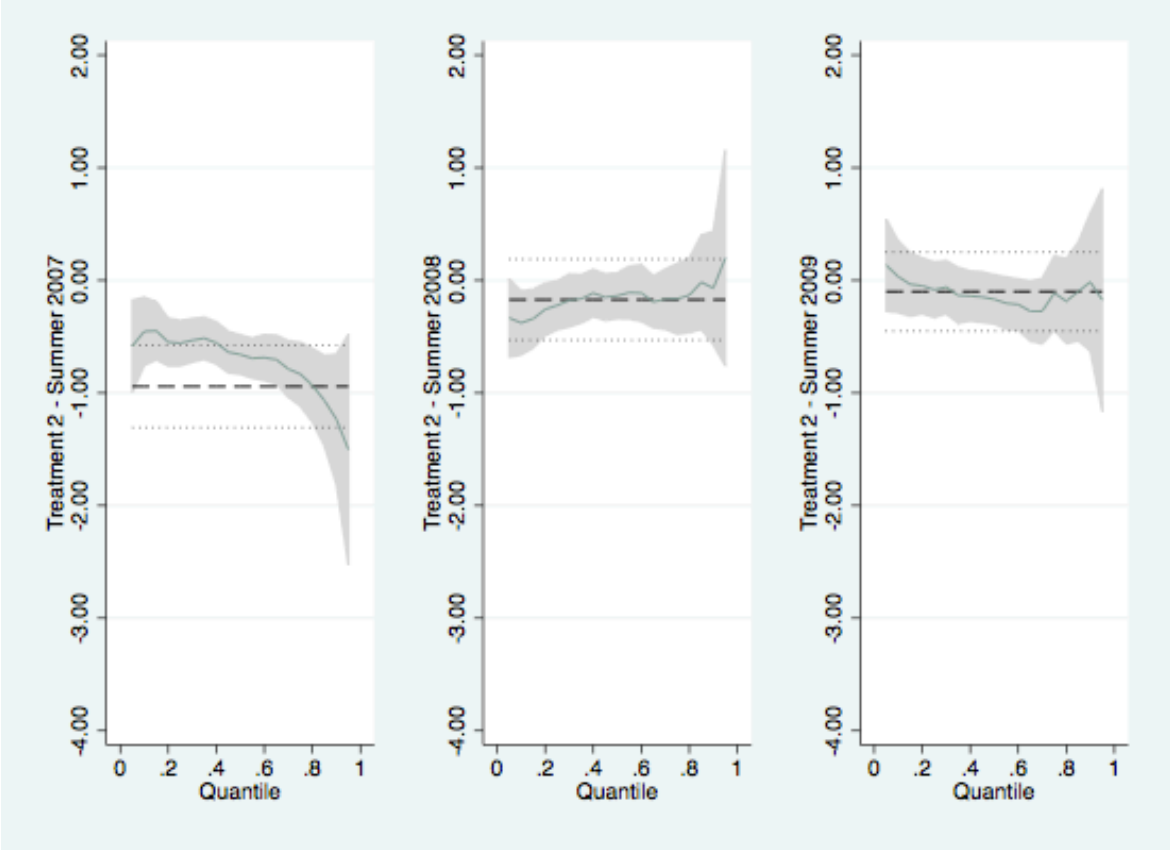
Note: All water consumption variables are in thousands of gallons.
Robust standard errors in parentheses
** p<0.01, * p<0.05

Figure 1
Quantile Treatment Effects for Treatment 1: Pure Information
(summer 2007, summer 2008, summer 2009)



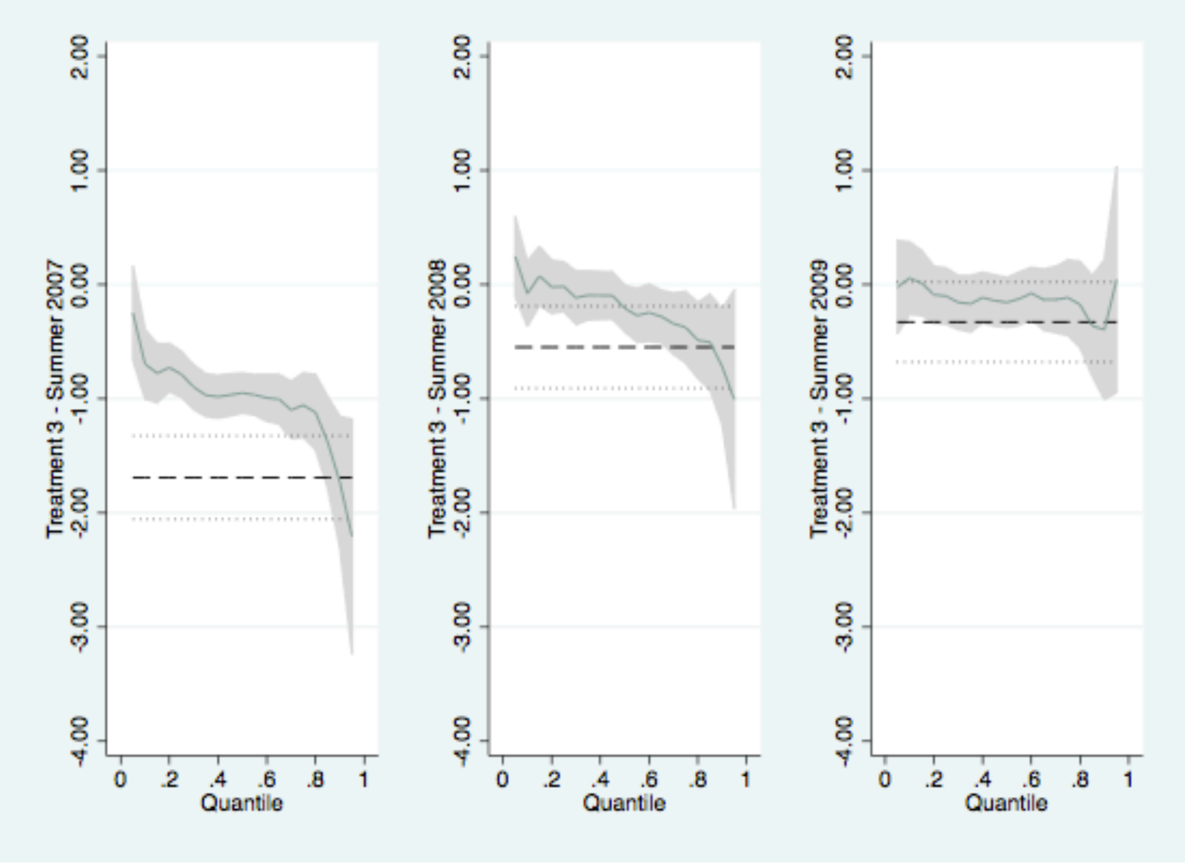
Note: Graphs plot quantile estimates and mean treatment effects across water use distribution. Dashed line depicts the Average Treatment Effect in a linear regression (OLS) framework, while dotted line represents its confidence interval. Solid line depicts the Quantile Treatment Effect and the shadowed area represents its confidence interval.

Figure 2
Quantile Treatment Effects for Treatment 2: Weak Social Norm
(summer 2007, summer 2008, summer 2009)



Note: Graphs plot quantile estimates and mean treatment effects across water use distribution. Dashed line depicts the Average Treatment Effect in a linear regression (OLS) framework, while dotted line represents its confidence interval. Solid line depicts the Quantile Treatment Effect and the shadowed area represents its confidence interval.

Figure 3
Quantile Treatment Effects for Treatment 3: Strong Social Norm
(summer 2007, summer 2008, summer 2009)



Note: Graphs plot quantile estimates and mean treatment effects across water use distribution. Dashed line depicts the Average Treatment Effect in a linear regression (OLS) framework, while dotted line represents its confidence interval. Solid line depicts the Quantile Treatment Effect and the shadowed area represents its confidence interval.

Appendix 1
Test of Zero CATE & Constant CATE

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	Zero CATE					Constant CATE				
	Chi-Sq	dof	*p-val*	Normal	*p-val*	Chi-Sq	dof	*p-val*	Normal	*p-val*
Top Down Selection of Covariates										
Summer 2007										
Treatment 1	36.47	13	0.001	4.60	0.000	25.44	12	0.013	2.74	0.003
Treatment 2	43.79	12	0.000	6.49	0.000	19.49	11	0.053	1.81	0.035
Treatment 3	230.32	15	0.000	39.31	0.000	175.10	14	0.000	30.44	0.000
Summer 2008										
Treatment 1	52.53	15	0.000	6.85	0.000	52.07	14	0.000	7.20	0.000
Treatment 2	14.21	11	0.222	0.68	0.247	12.91	10	0.229	0.65	0.258
Treatment 3	191.62	13	0.000	35.03	0.000	176.44	12	0.000	33.57	0.000
Summer 2009										
Treatment 1	67.98	11	0.000	12.15	0.000	67.16	10	0.000	12.78	0.000
Treatment 2	21.48	8	0.006	3.37	0.000	20.84	7	0.004	3.70	0.000
Treatment 3	189.92	11	0.000	38.15	0.000	185.52	10	0.000	39.25	0.000
Bottom Up Selection of Covariates										
Summer 2007										
Treatment 1	34.76	9	0.000	6.07	0.000	25.04	8	0.002	4.26	0.000
Treatment 2	37.57	10	0.000	6.16	0.000	13.06	9	0.160	0.96	0.170
Treatment 3	203.97	12	0.000	39.19	0.000	156.30	11	0.000	30.98	0.000
Summer 2008										
Treatment 1	48.47	10	0.000	8.60	0.000	48.12	9	0.000	9.22	0.000
Treatment 2	14.68	9	0.100	1.34	0.090	11.40	8	0.180	0.85	0.197
Treatment 3	126.37	14	0.000	21.24	0.000	120.26	13	0.000	21.04	0.000
Summer 2009										
Treatment 1	67.07	9	0.000	13.69	0.000	66.17	8	0.000	14.54	0.000
Treatment 2	20.72	8	0.008	3.18	0.001	19.73	7	0.006	3.40	0.000
Treatment 3	54.34	9	0.000	10.69	0.000	54.17	8	0.000	11.54	0.000
All Covariates										
Summer 2007										
Treatment 1	35.02	20	0.020	2.37	0.009	31.87	19	0.032	2.09	0.018
Treatment 2	23.43	20	0.268	0.54	0.294	19.90	19	0.401	0.15	0.442
Treatment 3	78.62	20	0.000	9.27	0.000	64.76	19	0.000	7.42	0.000
Summer 2008										
Treatment 1	50.30	20	0.000	4.79	0.000	49.94	19	0.000	5.02	0.000
Treatment 2	13.10	20	0.873	-1.09	0.862	12.68	19	0.855	-1.03	0.153
Treatment 3	71.26	20	0.000	8.10	0.000	66.81	19	0.000	7.76	0.000
Summer 2009										
Treatment 1	82.75	20	0.000	9.92	0.000	82.74	19	0.000	10.34	0.000
Treatment 2	14.71	20	0.793	-0.84	0.799	14.66	19	0.744	-0.70	0.241
Treatment 3	59.21	20	0.000	6.20	0.000	56.61	19	0.000	6.10	0.000

Zero CATE. H_0: Average Effect for subpopulation with covariates value X is equal to zero for all X. H_1: Average Effect for subpopulation with covariates value X is different from zero for some X.

Constant CATE. H_0: Average Effect for subpopulation with covariates value X is equal to ATE for all X. H_1: Average Effect for subpopulation with covariates value X is different from ATE for some X.

Note: For the zero and constant conditional average treatment effect test, the chi-sq column is equal to the square root of 2K times the normal column plus K, where K is the degrees of freedom. For the column with the zero average treatment effect results, the chi-sq column is equal to the square of the normal column.

Appendix 2
Rank Preservation Test
Treatment – Control Difference at Quantiles of the Outcome Distribution

Variable	(1)	(2)	(3)	(4)
	0 - 25th Percentile	25 - 50th Percentile	50 - 75th Percentile	75 - 100th Percentile
Treatment 1 (Technical Advice)				
Water Use Jun-Nov 2006 1/	-0.0880	-0.8591 **	-0.9299 **	-0.2527
Water Use Apr-May 2007 1/	-0.0161	-0.1419	-0.1908	-0.4425
House's Fair Market Value	-991.46	-641.91	-175.90	1,380.48
Age of House	0.1549	-0.4764 *	0.4511 *	-0.1842
% Owner Occupiers	-0.0031	0.0023	-0.0027	0.0012
% Population 25 years ≥ Bachelor 2/	-0.0028	0.0031	-0.0004	-0.0021
% of Households White 2/	-0.0025	0.0000	-0.0007	0.0014
Treatment 2 (Weak Social Norm)				
Water Use Jun-Nov 2006 1/	-0.0257	-0.7229 **	-0.7690 *	-2.4877 **
Water Use Apr-May 2007 1/	-0.0222	-0.2729 ***	-0.1059	-0.8508 **
House's Fair Market Value	-5,560.58 **	-1,278.2	-10,617.2 ***	-2,128.8
Age of House	-0.4089	0.2035	0.1609	-0.0550
% Owner Occupiers	0.0010	0.0082	0.0011	0.0194 ***
% Population 25 years ≥ Bachelor 2/	-0.0005	-0.0018	-0.0021	-0.0004
% of Households White 2/	-0.0002	-0.0027	-0.0018	0.0002
Treatment 3 (Strong Social Norm)				
Water Use Jun-Nov 2006 1/	0.047	-1.4748 ***	-2.6530 ***	-2.3427 **
Water Use Apr-May 2007 1/	-0.161	-0.4504 ***	-0.8378 ***	-0.5060
House's Fair Market Value	1,170	-4,894.4 **	-6,122.8 **	-14,537.2 ***
Age of House	-0.100	0.1153	0.3231	0.0554
% Owner Occupiers	0.001	-0.0007	0.0119	0.0232 ***
% Population 25 years ≥ Bachelor 2/	0.000	-0.0037	-0.0027	-0.0009
% of Households White 2/	0.000	-0.0035	0.0028	-0.0037

*** $p<0.01$, ** $p<0.05$, * $p<0.10$

1/ In thousands of gallons

2/ At block group level

Appendix 3
Mean on Summer 2006 and F-Test within subgroups
(in Thousand of gallons)

Subgroup		Treat 1	Treat 2	Treat 3	Control	F-Statistic	p-value
Previous water use (June - Nov 2006)	Below Median	22.46	22.37	22.45	22.42	0.350	0.786
	Above Median	57.48	57.60	57.36	57.40	0.060	0.981
Previous water use (April - May 2007)	Below Median	27.91	27.24	27.32	27.51	2.040	0.106
	Above Median	53.52	53.79	53.55	53.50	0.090	0.964
Fair Market Value	Below Median	31.04	30.15	30.37	30.36	2.700	0.044
	Above Median	48.14	48.59	48.92	48.54	0.460	0.712
Ownership Status	Renter	38.75	38.65	38.98	38.66	0.060	0.982
	Owner	39.75	39.42	39.71	39.61	0.240	0.872
Age of Home	Below Median	41.79	42.05	42.35	41.86	0.420	0.739
	Above Median	37.35	36.42	36.83	36.99	1.300	0.274
% with Higher Degree	Below Median	33.54	33.59	33.66	33.73	0.140	0.934
	Above Median	44.56	44.03	44.51	44.35	0.270	0.850
% White	Below Median	33.22	33.25	33.19	33.45	0.350	0.787
	Above Median	44.88	44.36	44.84	44.60	0.290	0.834

1/ At block group level

Appendix 4
Subgroup Analysis All Together

	(1)	(2)	(3)
	Summer	Summer	Summer
	2007	2008	2009
Treatment 1 (Technical Advice)	0.158	-0.0871	0.263
	(0.893)	(0.575)	(0.577)
Treatment 2 (Weak Social Norm)	-0.154	-0.173	0.613
	(0.764)	(0.599)	(0.579)
Treatment 3 (Strong Social Norm)	1.671*	0.193	0.580
	(0.703)	(0.559)	(0.549)
High - Previous water use (June - Nov 2006) 1/	15.51**	6.528**	8.512**
	(0.184)	(0.155)	(0.153)
High - Previous water use (April - May 2007) 1/	16.45**	9.324**	9.367**
	(0.199)	(0.163)	(0.161)
High - Fair Market Value 1/	7.997**	2.839**	5.127**
	(0.214)	(0.171)	(0.170)
Owners/Renters 2/	0.144	-1.318**	0.513*
	(0.334)	(0.248)	(0.230)
High - Age of Home 1/	-0.0427	0.168	-0.920**
	(0.205)	(0.153)	(0.156)
High - % White 1/	1.659**	-0.596**	0.422*
	(0.227)	(0.179)	(0.182)
High - % with higher degree 1/	0.492*	0.117	-0.0736
	(0.223)	(0.175)	(0.182)
Treat 1* High Water use June - Nov 2006	0.650	0.106	0.132
	(0.512)	(0.389)	(0.399)
Treat 2* High Water use June - Nov 2006	-0.480	-0.276	0.203
	(0.512)	(0.426)	(0.402)
Treat 3* High Water use June - Nov 2006	-1.131*	-0.319	-0.515
	(0.476)	(0.390)	(0.408)
Treat 1* High Water use April - May 2006	0.00835	0.133	0.595
	(0.564)	(0.414)	(0.426)
Treat 2* High Water use April - May 2006	-0.265	0.151	0.0571
	(0.541)	(0.457)	(0.424)
Treat 3* High Water use April - May 2006	-1.341**	-0.549	0.239
	(0.510)	(0.404)	(0.413)
Treat 1 * High Fair Market Value	-1.141	-0.667	-0.836
	(0.589)	(0.433)	(0.449)
Treat 2 * High Fair Market Value	-0.0950	0.193	-0.178
	(0.573)	(0.430)	(0.430)
Treat 3 * High Fair Market Value	-1.349*	-0.661	-1.158**
	(0.526)	(0.489)	(0.431)
Treat 1 * Owner	-0.828	0.312	-0.478
	(1.054)	(0.600)	(0.644)
Treat 2 * Owner	0.0773	-0.235	-0.418
	(0.777)	(0.607)	(0.567)
Treat 3 * Owner	-1.617*	-0.0247	-0.237
	(0.777)	(0.569)	(0.553)
Treat 1 * High Age of Home	-0.370	-0.316	0.151
	(0.577)	(0.415)	(0.422)

Cont.

	(1) Summer 2007	(2) Summer 2008	(3) Summer 2009
Treat 2 * High Age of Home	-0.399	0.0375	-0.672
	(0.511)	(0.433)	(0.405)
Treat 3 * High Age of Home	-0.501	0.203	-0.387
	(0.491)	(0.407)	(0.381)
Treat 1 * High % White	1.018	0.139	0.463
	(0.639)	(0.485)	(0.477)
Treat 2 * High % White	-0.0607	0.474	0.632
	(0.633)	(0.487)	(0.484)
Treat 3 * High % White	0.550	0.0398	0.247
	(0.574)	(0.470)	(0.465)
Treat 1 * High % with higher degree	0.840	0.442	0.400
	(0.596)	(0.448)	(0.453)
Treat 2 * High % with higher degree	-0.101	0.0141	-0.495
	(0.641)	(0.478)	(0.467)
Treat 3 * High % with higher degree	-0.267	-0.204	0.224
	(0.575)	(0.444)	(0.445)
Constant	16.10**	18.06**	16.32**
	(0.311)	(0.238)	(0.231)
Observations	94,411	94,411	94,411
R-squared	0.300	0.138	0.192
P-value Test Equal Impact in All Subgroups in Treat1	0.211	0.828	0.283
P-value Test Equal Impact in All Subgroups in Treat2	0.888	0.906	0.632
P-value Test Equal Impact in All Subgroups in Treat3	0.000	0.152	0.141
P-value Test Equal Impact in All Subgroups in All Treatments	0.000	0.545	0.253

1/ Dummy variables. High subgroup defined above median.

2/ Dummy variable. Owners (=1). Renters (=0).

Robust standard errors in parentheses

** p<0.01, * p<0.05

Appendix 5
Subgroup Analysis for Water Consumption in summer 2008

	Dependent Variable: Summer 2008						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Previous water use (June - Nov 2006)	Previous water use (April - May 2007)	Fair Market Value	Owners / Renters 1/	Age of Home	% White	% with higher degree
Treatment 1	0.0270	-0.00534	0.242	-0.495	0.149	-0.0307	-0.0976
(Technical Advice)	(0.158)	(0.159)	(0.210)	(0.537)	(0.291)	(0.263)	(0.254)
Treatment 2	0.0152	-0.0184	-0.270	0.239	-0.0618	-0.306	-0.304
(Weak Social Norm)	(0.169)	(0.163)	(0.212)	(0.562)	(0.339)	(0.270)	(0.268)
Treatment 3	-0.0352	-0.174	-0.0202	-0.00421	-0.774**	-0.291	-0.196
(Strong Social Norm)	(0.152)	(0.160)	(0.203)	(0.520)	(0.259)	(0.267)	(0.262)
Subgroup var. (high = 1)	12.10**	13.23**	5.774**	-0.306	-1.307**	2.647**	2.782**
	(0.146)	(0.150)	(0.149)	(0.242)	(0.150)	(0.158)	(0.158)
Treat1*high subgrop var.	-0.0691	0.160	-0.444	0.606	-0.214	0.135	0.262
	(0.383)	(0.395)	(0.393)	(0.577)	(0.397)	(0.416)	(0.416)
Treat2*high subgrop var.	-0.206	-0.148	0.266	-0.488	-0.245	0.185	0.178
	(0.424)	(0.438)	(0.428)	(0.609)	(0.429)	(0.454)	(0.454)
Treat3*high subgrop var.	-1.047**	-0.924*	-0.931*	-0.602	0.536	-0.348	-0.517
	(0.357)	(0.362)	(0.368)	(0.556)	(0.370)	(0.389)	(0.389)
Constant	19.60**	19.40**	22.60**	25.75**	26.13**	24.25**	24.18**
	(0.0597)	(0.0648)	(0.0795)	(0.229)	(0.110)	(0.0985)	(0.100)
Observations	102,887	102,887	102,871	102,869	102,461	94,833	94,833
R-squared	0.091	0.110	0.021	0.000	0.001	0.004	0.005
p-value equal impact T1	0.857	0.686	0.258	0.294	0.590	0.745	0.528
p-value equal impact T2	0.628	0.735	0.534	0.423	0.567	0.684	0.695
p-value equal impact T3	0.003	0.011	0.011	0.279	0.148	0.371	0.184

Note: All water consumption variables are in thousands of gallons.
1/ In the case of Owners (=1) / Renter (=0), interaction terms are for owner group rather than high group.
Robust standard errors in parentheses
** p<0.01, * p<0.05

Appendix 6
Subgroup Analysis for Water Consumption in summer 2009

	Dependent Variable: Summer 2009						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Previous water use (June - Nov 2006)	Previous water use (April - May 2007)	Fair Market Value	Owners / Renters 1/	Age of Home	% White	% with higher degree
Treatment 1	0.127	0.112	0.520*	0.568	0.387	0.0644	0.0856
(Technical Advice)	(0.164)	(0.173)	(0.212)	(0.595)	(0.318)	(0.261)	(0.266)
Treatment 2	-0.127	-0.121	-0.126	0.465	0.494	-0.169	-0.0260
(Weak Social Norm)	(0.165)	(0.173)	(0.204)	(0.534)	(0.340)	(0.267)	(0.266)
Treatment 3	-0.0342	-0.264	0.164	0.112	-0.201	-0.252	-0.180
(Strong Social Norm)	(0.166)	(0.166)	(0.207)	(0.511)	(0.308)	(0.259)	(0.260)
Subgroup var. (high = 1)	15.30**	15.35**	9.654**	2.044**	-3.321**	5.022**	4.526**
	(0.150)	(0.156)	(0.155)	(0.228)	(0.158)	(0.164)	(0.164)
Treat1*high subgrop var.	0.417	0.674	-0.303	-0.244	-0.00330	0.375	0.323
	(0.403)	(0.419)	(0.417)	(0.637)	(0.426)	(0.437)	(0.437)
Treat2*high subgrop var.	0.392	0.377	0.265	-0.587	-1.075*	-0.0320	-0.330
	(0.408)	(0.422)	(0.418)	(0.583)	(0.426)	(0.431)	(0.432)
Treat3*high subgrop var.	-0.529	-0.197	-0.726	-0.389	-0.0376	-0.0566	-0.151
	(0.375)	(0.386)	(0.391)	(0.555)	(0.399)	(0.409)	(0.410)
Constant	19.96**	20.36**	22.58**	25.69**	29.06**	24.79**	25.04**
	(0.0636)	(0.0663)	(0.0776)	(0.211)	(0.120)	(0.101)	(0.103)
Observations	102,887	102,887	102,871	102,869	102,461	94,833	94,833
R-squared	0.136	0.137	0.053	0.001	0.007	0.015	0.012
p-value equal impact T1	0.301	0.108	0.468	0.701	0.994	0.390	0.460
p-value equal impact T2	0.337	0.373	0.527	0.314	0.012	0.941	0.445
p-value equal impact T3	0.159	0.610	0.063	0.483	0.925	0.890	0.713

Note: All water consumption variables are in thousands of gallons.

1/ In the case of Owners (=1) / Renter (=0), interaction terms are for owner group rather than high group.

Robust standard errors in parentheses

** p<0.01, * p<0.05

Appendix 7
Block Price Threshold Regressions by Treatment: Measure 1
(Dummy above threshold)

	(1)	(2)	(3)	(4)	(5)
	9,000	16,000	9,000	16,000	9,000
	bw=500		bw=1,000		bw=2,000
Treatment 1	-0.152	2.026	-0.390	0.956	-0.293
(Technical Advice)	(0.820)	(2.585)	(0.534)	(1.763)	(0.330)
Treatment 2	-1.093	-1.086	-1.275*	-0.911	-0.562
(Weak Social Norm)	(0.811)	(2.750)	(0.521)	(1.791)	(0.332)
Treatment 3	-1.473*	-3.600	-1.163*	-4.378**	-1.119**
(Strong Social Norm)	(0.641)	(2.274)	(0.480)	(1.631)	(0.346)
Dummy Above Bandwidth (bw)	0.921	1.335	2.560**	2.179*	5.865**
	(0.472)	(1.445)	(0.326)	(0.975)	(0.240)
Treat1*Above Bandwidth	0.878	-1.566	1.030	3.413	-0.0295
	(1.182)	(3.691)	(0.835)	(3.523)	(0.597)
Treat2*Above Bandwidth	1.849	-4.285	1.215	-2.140	-0.227
	(1.368)	(3.440)	(0.905)	(2.445)	(0.605)
Treat3*Above Bandwidth	2.409*	-4.336	1.345	-0.843	0.206
	(1.084)	(3.320)	(0.841)	(2.437)	(0.634)
Fair Market Value	1.98e-05**	3.00e-05**	1.69e-05**	3.20e-05**	1.63e-05**
	(2.65e-06)	(4.39e-06)	(3.19e-06)	(5.11e-06)	(1.86e-06)
Age of Home	0.0223	0.132*	0.00432	0.0813	0.00374
	(0.0169)	(0.0642)	(0.0120)	(0.0437)	(0.00804)
Ownership status	1.292*	6.519**	0.892*	5.298**	0.728**
	(0.548)	(1.927)	(0.409)	(1.364)	(0.277)
% White	3.617*	1.780	3.928**	11.48*	4.566**
	(1.723)	(7.985)	(1.205)	(5.237)	(0.825)
% with Higher Degree	-5.993**	-13.64*	-5.481**	-12.72**	-6.109**
	(1.652)	(6.165)	(1.459)	(4.540)	(0.939)
Constant	28.16**	45.92**	28.38**	37.35**	26.76**
	(1.365)	(6.098)	(0.959)	(4.036)	(0.667)
Observations	6,617	1,715	13,338	3,402	27,335
R-squared	0.027	0.049	0.031	0.060	0.057

Robust standard errors in parentheses
** p<0.01, * p<0.05

Cont.

	(6)	(7)	(8)	(9)	(10)
	16,000	9,000	16,000	9,000	16,000
	bw=2,000	bw=3,000		bw=4,000	
Treatment 1	-0.988	-0.467	0.197	-0.504*	0.230
(Technical Advice)	(1.138)	(0.240)	(0.917)	(0.196)	(0.917)
Treatment 2	-1.799	-0.628*	-1.314	-0.609**	-1.289
(Weak Social Norm)	(1.199)	(0.272)	(0.915)	(0.218)	(0.915)
Treatment 3	-4.798**	-0.968**	-4.192**	-0.919**	-4.148**
(Strong Social Norm)	(1.069)	(0.257)	(0.818)	(0.203)	(0.819)
Dummy Above Bandwidth (bw)	4.281**	8.763**	7.349**	11.65**	3.325**
	(0.726)	(0.206)	(0.595)	(0.187)	(0.476)
Treat1*Above Bandwidth	2.457	0.177	0.709	0.0943	-0.0953
	(2.200)	(0.511)	(1.786)	(0.460)	(1.333)
Treat2*Above Bandwidth	-0.930	-0.283	-0.819	-0.523	-0.631
	(1.827)	(0.521)	(1.472)	(0.474)	(1.264)
Treat3*Above Bandwidth	0.182	-0.276	1.582	-0.387	1.841
	(1.779)	(0.538)	(1.529)	(0.486)	(1.195)
Fair Market Value	2.68e-05**	1.77e-05**	2.71e-05**	1.95e-05**	3.03e-05**
	(3.00e-06)	(1.48e-06)	(2.29e-06)	(1.28e-06)	(1.86e-06)
Age of Home	0.0255	0.00843	0.0383	0.0114*	0.0308
	(0.0278)	(0.00669)	(0.0230)	(0.00538)	(0.0190)
Ownership status	3.145**	0.423	2.357**	0.190	1.971**
	(0.973)	(0.227)	(0.776)	(0.188)	(0.644)
% White	5.956	4.078**	5.791	3.884**	7.289**
	(3.670)	(0.645)	(3.055)	(0.528)	(2.467)
% with Higher Degree	-4.854	-5.897**	-6.135*	-5.066**	-5.192*
	(3.136)	(0.745)	(2.529)	(0.610)	(2.075)
Constant	40.14**	25.18**	39.48**	22.69**	36.97**
	(2.842)	(0.520)	(2.291)	(0.429)	(1.850)
Observations	6,866	42,640	10,619	59,298	14,772
R-squared	0.053	0.097	0.066	0.146	0.054

Robust standard errors in parentheses
** p<0.01, * p<0.05

Appendix 8
Block Price Threshold Regressions by Treatment: Measure 2
(Continuous Variable, Difference with respect to Threshold)

	(1)	(2)	(3)	(4)	(5)
	9,000	16,000	9,000	16,000	9,000
	bw=500		bw=1,000		bw=2,000
Treatment 1 (Technical Advice)	0.641	-2.078	0.522	-0.213	0.736
	(1.065)	(3.384)	(0.820)	(2.753)	(0.579)
Treatment 2 (Weak Social Norm)	0.322	-4.030	0.210	-4.593*	-0.397
	(1.343)	(2.499)	(0.962)	(2.108)	(0.644)
Treatment 3 (Strong Social Norm)	0.513	-5.845	0.000839	-5.974*	-0.0443
	(0.881)	(3.069)	(0.745)	(2.360)	(0.552)
Diff. Threshold – Consump. (Diff.)	-0.974	-1.681	-0.0718	-1.032	-0.0863
	(0.927)	(2.656)	(0.552)	(1.627)	(0.373)
Dummy Above Bandwidth	0.656	2.207	0.422*	1.028*	0.715**
	(0.537)	(1.658)	(0.173)	(0.497)	(0.0545)
Treat1*Dummy Above Bandwidth	0.598	-2.948	0.330	-0.439	0.180
	(0.796)	(2.696)	(0.322)	(1.056)	(0.110)
Treat2*Dummy Above Bandwidth	0.763	-1.978	0.500	-1.270	-0.0148
	(0.944)	(2.469)	(0.355)	(0.921)	(0.124)
Treat3*Dummy Above Bandwidth	1.217	-0.544	0.363	-0.0999	0.197
	(0.648)	(2.291)	(0.312)	(0.983)	(0.111)
Dummy Above BW * Diff.	1.290	-3.652	0.729*	-0.874	0.134
	(0.904)	(2.836)	(0.287)	(0.841)	(0.110)
Treat 1 * Dummy Above BW * Diff.	-0.356	7.621	-0.281	3.612	-0.595*
	(2.149)	(6.848)	(0.761)	(3.425)	(0.258)
Treat 2 * Dummy Above BW * Diff.	-0.302	-1.061	-0.840	2.567	-0.242
	(2.695)	(5.581)	(0.846)	(1.936)	(0.274)
Treat 3 * Dummy Above BW * Diff.	-0.633	0.305	-0.430	1.499	-0.587*
	(2.046)	(6.214)	(0.767)	(2.184)	(0.256)
Fair Market Value	1.98e-05**	3.00e-05**	1.68e-05**	3.17e-05**	1.57e-05**
	(2.64e-06)	(4.38e-06)	(3.20e-06)	(5.13e-06)	(1.82e-06)
Age of Home	0.0220	0.132*	0.00398	0.0770	0.00252
	(0.0169)	(0.0645)	(0.0120)	(0.0442)	(0.00800)
Ownership status	1.320*	6.415**	0.908*	5.374**	0.773**
	(0.547)	(1.929)	(0.409)	(1.364)	(0.275)
% White	3.694*	1.454	3.953**	10.93*	4.385**
	(1.719)	(7.993)	(1.201)	(5.301)	(0.819)
% with Higher Degree	-6.088**	-13.42*	-5.642**	-12.87**	-6.118**
	(1.644)	(6.168)	(1.461)	(4.558)	(0.926)
Constant	29.12**	49.68**	29.53**	40.88**	30.43**
	(1.572)	(6.567)	(1.054)	(4.294)	(0.710)
Observations	6,617	1,715	13,338	3,402	27,335
R-squared	0.029	0.051	0.035	0.062	0.070

Robust standard errors in parentheses
** p<0.01, * p<0.05

Cont.

	(6)	(7)	(8)	(9)	(10)
	16,000	9,000	16,000	9,000	16,000
	bw=2,000	bw=3,000		bw=4,000	
Treatment 1 (Technical Advice)	4.190*	0.221	1.596	0.0752	0.320
	(2.116)	(0.474)	(1.779)	(0.401)	(0.897)
Treatment 2 (Weak Social Norm)	-2.246	-0.326	-2.298	-0.398	-1.773*
	(1.699)	(0.500)	(1.376)	(0.426)	(0.800)
Treatment 3 (Strong Social Norm)	-5.146**	-0.604	-6.831**	-0.745	-2.676**
	(1.697)	(0.470)	(1.402)	(0.404)	(0.788)
Diff. Threshold – Consump. (Diff.)	-0.601	-0.0170	-1.068	-0.134	-0.896
	(1.155)	(0.301)	(0.939)	(0.262)	(0.690)
Dummy Above Bandwidth	0.454*	0.725**	0.780**	0.744**	0.737**
	(0.196)	(0.0283)	(0.0906)	(0.0170)	(0.0847)
Treat1*Dummy Above Bandwidth	1.037*	0.0776	0.136	0.0528	-0.0118
	(0.413)	(0.0592)	(0.236)	(0.0376)	(0.156)
Treat2*Dummy Above Bandwidth	-0.0234	0.0253	-0.0996	0.0156	-0.0394
	(0.359)	(0.0647)	(0.188)	(0.0398)	(0.143)
Treat3*Dummy Above Bandwidth	0.0613	0.0402	-0.330	0.0172	0.137
	(0.349)	(0.0603)	(0.189)	(0.0369)	(0.135)
Dummy Above BW * Diff.	0.331	0.0572	-0.193	0.0278	-0.0994
	(0.324)	(0.0591)	(0.171)	(0.0392)	(0.0876)
Treat 1 * Dummy Above BW * Diff.	-1.968*	-0.230	-0.413	-0.161	0.108
	(0.815)	(0.149)	(0.476)	(0.0958)	(0.150)
Treat 2 * Dummy Above BW * Diff.	0.0794	-0.193	0.172	-0.157	0.0897
	(0.845)	(0.150)	(0.401)	(0.103)	(0.144)
Treat 3 * Dummy Above BW * Diff.	0.307	-0.234	1.261**	-0.143	-0.107
	(0.778)	(0.150)	(0.446)	(0.0978)	(0.139)
Fair Market Value	2.68e-05**	1.62e-05**	2.64e-05**	1.68e-05**	2.59e-05**
	(3.01e-06)	(1.40e-06)	(2.30e-06)	(1.18e-06)	(1.86e-06)
Age of Home	0.0261	0.00688	0.0380	0.00788	0.0310
	(0.0278)	(0.00661)	(0.0228)	(0.00525)	(0.0184)
Ownership status	3.121**	0.469*	2.387**	0.293	1.985**
	(0.974)	(0.224)	(0.776)	(0.184)	(0.634)
% White	5.662	3.853**	5.607	3.282**	5.707*
	(3.664)	(0.635)	(3.048)	(0.513)	(2.417)
% with Higher Degree	-5.019	-6.033**	-6.538**	-5.311**	-5.614**
	(3.137)	(0.722)	(2.517)	(0.578)	(2.043)
Constant	42.78**	30.94**	45.66**	31.00**	45.15**
	(3.026)	(0.556)	(2.398)	(0.454)	(1.920)
Observations	6,866	42,640	10,619	59,298	14,772
R-squared	0.057	0.123	0.077	0.192	0.106

Robust standard errors in parentheses
** p<0.01, * p<0.05

Appendix 9
Block Price Threshold Regressions by Treatment: Measure 3
(Dummy above threshold, for each month, rather than aggregated)

	(1)	(2)	(3)	(4)	(5)	(6)
	9,000	16,000	9,000	16,000	9,000	16,000
	bw=2,000		bw=3,000		bw=4,000	
Treatment 1	-0.622	-4.949	-0.0940	-0.655	-0.648**	0.733
(Technical Advice)	(0.659)	(4.075)	(0.377)	(2.676)	(0.245)	(3.120)
Treatment 2	0.0479	1.158	-0.244	0.310	-0.705**	-1.069
(Weak Social Norm)	(0.643)	(6.289)	(0.360)	(2.812)	(0.246)	(1.688)
Treatment 3	-0.903	-3.634	-1.145**	-4.102	-0.630*	-2.743
(Strong Social Norm)	(0.573)	(4.997)	(0.328)	(2.682)	(0.266)	(1.702)
Dummy Above Bandwidth	3.459**	1.980	6.880**	9.576*	9.854**	7.847**
	(0.891)	(8.776)	(0.562)	(3.883)	(0.468)	(2.657)
Treat1*Above Bandwidth	5.120	----	-0.589	-2.066	1.905	-2.031
	(3.164)	----	(1.567)	(6.462)	(1.438)	(6.307)
Treat2*Above Bandwidth	5.809	-3.616	1.307	-1.315	1.255	9.506
	(4.157)	(12.50)	(1.606)	(6.180)	(1.206)	(11.39)
Treat3*Above Bandwidth	2.483	----	2.134	-2.212	0.870	-7.241
	(3.216)	----	(1.676)	(11.00)	(1.271)	(6.219)
Fair Market Value	1.24e-05**	2.51e-05*	1.75e-05**	1.96e-05**	1.77e-05**	2.46e-05**
	(2.91e-06)	(9.89e-06)	(3.06e-06)	(5.20e-06)	(1.77e-06)	(4.24e-06)
Age of Home	0.0128	0.183	0.0101	0.158	0.0228**	0.169**
	(0.0156)	(0.124)	(0.00903)	(0.0867)	(0.00692)	(0.0504)
Ownership status	1.049	4.221	0.635	0.409	0.490*	0.834
	(0.541)	(4.286)	(0.342)	(2.596)	(0.245)	(1.538)
% White	0.718	16.66	0.209	-3.841	1.105	-6.324
	(1.616)	(16.25)	(0.955)	(8.637)	(0.673)	(5.704)
% with Higher Degree	-1.031	-12.52	-1.644	4.655	-2.082**	10.50
	(1.639)	(12.97)	(1.078)	(7.429)	(0.732)	(5.425)
Constant	28.62**	34.95**	27.21**	43.47**	24.69**	39.39**
	(1.398)	(12.16)	(0.800)	(7.303)	(0.563)	(4.728)
Observations	4,056	193	12,065	637	24,888	1,525
R-squared	0.022	0.069	0.041	0.044	0.053	0.043

Robust standard errors in parentheses
** p<0.01, * p<0.05

CHAPTER II:
COMPARING EXPERIMENTAL AND NON-EXPERIMENTAL EVALUATION DESIGNS USING A LARGE-SCALE RANDOMIZED EXPERIMENT

1. Introduction

Socioeconomic policies and programs are rarely implemented within a randomized design. Thus researchers aiming to estimate policy and program impacts must depend on non-experimental, observational designs. Non-experimental designs have experienced important advances over the last two decades, including innovations in panel data analysis, matching methods, inverse probability weighting and trimming rules. There remains, however, continued debate about how well these non-experimental designs can uncover causal relationships (Smith and Todd, 2005). Although non-experimental designs are often more feasible and, sometimes, less costly than controlled experiments, their ability to uncover causal relationships rests on untestable assumptions about the differences between treatment and control groups.

Starting with Lalonde (1986), a small but growing number of social scientists have tried to use randomized experiments to validate non-experimental econometric designs. In these studies, called "within-study comparisons" or "design replication studies," researchers estimate a program's impact by using a randomized control group. Then they re-estimate the impact by using one or more nonrandomized comparison groups and econometric techniques to eliminate or mitigate observable and unobservable sources of bias. The results from the randomized experimental design provide a benchmark for evaluating the success with which non-experimental designs can recover causal treatment effects. If the non-experimental estimate is close to the experimental estimate, which is assumed to have high internal validity, the non-experimental design is labeled as "successful."

Most of design-replication studies have been completed in the context of welfare, job training, or employment services programs with voluntary program participation. However, socioeconomic policies and programs are frequently implemented or piloted in administrative units like towns, counties, or states. Thus to estimate impacts, economists typically look to neighboring administrative units for comparison groups and apply various econometric techniques to control for observable and unobservable sources of bias. See, for example, the seminal article by Card and Krueger (1994) who use Pennsylvania fast-food restaurants as controls for New Jersey fast-food restaurants in order to assess the impact of minimum wage laws. Other noteworthy examples are Abadie and Gardeazabal (2003), who estimate the effects of the terrorist conflict in the Spanish Basque County on GDP using other Spanish regions as a comparison group; Besley and Burgess (2004), who evaluate the effect of pro-workers' labor regulation on firm productivity in India by comparing firms in states that adopted these regulations with firms in states that did not; Davis (2004), who estimate the impact of pediatric leukemia risk on local housing values by comparing households in a county in Nevada with a nearby county acting as a control group, and; Galiani, Gertler and Schargrodsky (2005), who evaluate the effect of water service privatization on child mortality in Argentina by comparing households in municipalities with and without privatized service.

Design replication studies based on voluntary program participation has the objective of reducing the selection bias problem, i.e. observable and unobservable differences between treated and control units. However, programs implemented in administrative units with all individuals or houses treated (or untreated), are less likely to have the problem of selection bias. Selection bias only arises from the pre-treatment household decision of where to reside and not from the participation decision due to the treatment. Thus, it should be simpler to replicate experimental results using non-experimental techniques under no selection bias due to the treatment itself. Therefore, extending the design-replication approach in a context where the selection problem is less problematic would be fruitful and will provide insights about the strengths of non-experimental techniques.

In 2007, a large-scale randomized field experiment tested the effects of three messages on voluntary reductions in water consumption among households during a drought in a county in metropolitan Atlanta, Georgia (Ferraro and Price, forthcoming). The treatment consisted on sending the letter (rather than reading it), thus there is 100% compliance. In this study, we focus on two of the three treatments: (i) a social comparison message, and (ii) a technical information message. Each treatment group comprised approximately 12,000 households and the control group comprised approximately 72,000 households. The social comparison treatment included technical information that explained how households could reduce water use, social norm-based encouragement, and a social comparison in which own consumption was compared to median county consumption. This treatment had a large and statistically significant estimated treatment effect. The technical information treatment, which only instructed households on strategies to reduce water use, had a small and statistically insignificant impact.

We form a non-experimental comparison group of approximately 67,000 households from a neighboring county. The neighboring county experienced similar water pricing policies, water sources, weather patterns, state and metro regulatory environments and other regional confounding factors during the information campaign experiment. Participants do not self-select into the program, but they may have sorted themselves across counties based on characteristics that also affect water consumption. Our administrative data comprise monthly water use for 17 months, including pre and post experiment periods. By merging the treatment and non-experimental control group data with tax assessor and census data, we create a unique data set that also includes home characteristics and block group characteristics from the 2000 US census.

Using the non-experimental comparison group, we estimate treatment effects of the conservation information campaigns with popular econometric approaches, including panel data estimators, which have rarely been assessed in the context of a design-replication study, and inverse propensity score weighting and trimming rules, which have not yet been assessed. We use bootstrapping methods to estimate the distribution of the non-experimental treatment effect estimates. We then compare these treatment effect estimates to the estimate derived from the randomized experimental design. Assuming no randomization or general equilibrium (spillover) biases, which were not detected in Ferraro and Price (forthcoming), randomization of households into control and treatment groups, followed by a comparison of each group’s mean water consumption, provides an unbiased estimator of the average treatment effect.

We find that the selection of control units matters significantly in the performance of non-experimental econometric estimators. When matching methods are used to pre-process the data and make more similar the treated and control units’ pre-treatment average outcome trends and distribution of baseline covariates, simple panel data methods, with fixed household effects, generate estimates almost identical to the experimental estimates. Statistical inferences are also identical. However, using alternative designs (e.g., panel methods without pre-processing by matching, or matching followed by single difference and difference-in-differences estimators), the non-experimental estimators are inaccurate.

The next section summarizes the literature on the performance of non-experimental evaluation designs. Section 3 describes the study site and the experiment. Section 4 describes non-experimental designs used in this study. Section 5 describes the data. Section 6 presents the experimental benchmark. Section 7 presents the criteria used to evaluate the performance of non-experimental estimators. Section 8 shows non-experimental sample. Section 9 presents non-experimental estimates, and Section 10 presents distributions of the non-experimental treatment effect estimates.

2. The Performance of Non-Experimental Evaluation Designs

The literature on the performance of non-experimental designs can be classified into four different categories: computer simulations, meta-analyses, double randomized preference trials, and design-replication studies (Shadish et al., 2008 and Glazerman et al., 2003). *Computer simulation* studies produce controlled but artificial data varying key features that might affect outcome variables. These types of studies can provide accurate results, but the data are artificial (e.g. Drake, 1993).[14] *Meta-analyses* compare studies that examine the same (approximately) treatment but use different samples, different randomization designs or different methods (e.g. Greenberg et al., 2006, Glazerman et al., 2003). In a review, Shadish et al. (2008) note these meta-analyses yield mixed evidence on the ability of non-experimental designs to replicate the results of experimental designs, but the authors also note that meta-analyses are unable to control fully for differences across studies and thus their conclusions should be interpreted with caution.

Double randomized preference trials are experiments in which some subjects are randomly assigned to be in a randomized experiment, in which subjects are randomly assigned to one of multiple treatments, or a non-randomized experiment, in which subjects choose one of the same multiple treatments. Shadish et al. (2008) present the only example of this type of experiment, in which the treatments are training sessions in mathematics or vocabulary and the outcomes are exam performances.[15] Although double randomized preference trials have

[14] Drake (1993) generates artificial data and evaluates different sources of bias finding that (i) propensity score did not introduce additional bias, and (ii) omitted covariates introduces large bias.

[15] Shadish et al. (2008) found that OLS regressions reduced bias by 84-94%, while propensity score matching followed by simple tests of means reduced bias by an average of 74% (59-96%). No other designs were evaluated.

attractive characteristics from the perspective of comparing experimental and non-experimental designs, they are substantially more difficult to initiate than standard randomized controlled trials in realistic policy contexts (which are themselves difficult to initiate).

Design replication studies, which are also known as *within-study design*s, estimate a program's impact by using a randomized control group. Then they re-estimate the impact by using one or more nonrandomized comparison groups and econometric techniques to eliminate or mitigate observable and unobservable sources of bias. This type of study started with Lalonde (1986), Fraker and Maynard (1987) and Lalonde and Maynard (1987), who use data from the National Supported Work (NSW) randomized field experiment, and non-random comparison groups selected from national surveys such as the Current Population Survey (CPS) and the Panel Study of Income Dynamics (PSID). These authors argue that non-experimental methods cannot systematically recover experimental estimates of labor market program impacts.

Using data similar to the data used by Lalonde (1986), Dehejia and Wahba (1999, 2002) came to a more optimistic conclusion. They argue that Propensity Score Matching (PSM) can recover experimental estimators when the analyst has detailed information on pre-experiment outcomes and covariates. Smith and Todd (2005), however, suggest that the Dehejia and Wahba conclusion depends on their choice of a particular subsample of LaLonde (1986), rather than qualities of the PSM method. More generally, Smith and Todd (2005) state that the results of Dehejia and Wahba (1999, 2002) and Lalonde (1986) cannot be generalized because their data do not satisfy three criteria that Heckman and colleagues have argued are needed to draw unbiased (or small bias) inferences from matching estimators (Heckman et al., 1997, 1998a, 1998b): (i) participant and nonparticipant data come from the same sources, with similar measures of the outcome variable being most important; (ii) participants and nonparticipants share the same economic environment; and (iii) the data should contain a rich set of variables that affect both program participation and the outcome.[16]

[16] Heckman et al. (1997) specifically state that: "The major source of bias arising from the application of nonexperimental estimators to evaluate training programmes that is reported in LaLonde (1986) arises from the mismatch of questionnaires and labour markets across treatment group and comparison group members, and not because of the failure of econometric estimators to eliminate selection bias." (p. 608).

Outside of labor market programs, design-replication studies have been conducted on educational programs (Agodini and Dynarski, 2004; Hill et al., 2004; Wilde and Hollister, 2007), poverty reduction programs (Diaz and Handa, 2006; Handa and Maluccio, 2010), migration (McKenzie et al., 2010), and elections (Arceneaux et al., 2006). Most of these studies focus on the performance of PSM and none of them found systematic evidence that PSM can replicate experimental results (Agodini and Dynarski, 2004; Hill et al., 2004; Diaz and Handa, 2006; Handa and Maluccio, 2010; Wilde and Hollister, 2007). Overall, the non-experimental estimates are closer to the experimental estimates when Heckman et al. (1997, 1998a, 1998b) criteria hold. Specifically, the greater the use of pre-treatment variables (outcome and covariates) and the more similar the treated and control unit environments, the better are the results.

These previous studies focus on programs that involve voluntary self-selection into the program, whereas the program we study is more like a uniformly applied government policy (e.g., state-level minimum wage law increases) where participants do not choose to expose themselves to the treatment, hence selection bias due to the treatment is minimized. Moreover, in contrast to our study, none of previous design-replication studies have focused on environmental outcomes, information-based treatments, or programs with 100% compliance due to the fact that the treatment consists on sending the letters. Thus, our design replication study contributes to the design-replication literature in several and important ways.

With regard to the designs assessed in the literature, there are only a few design-replication studies that use panel data (Smith and Todd, 2005; Heckman et al., 1997; Heckman et al., 1998b). Results in Heckman et al. (1997 and 1998b) and Smith and Todd (2005), using data from the same job training program,[17] suggest that a difference-in-differences design, which helps eliminate individual, time-invariant unobserved heterogeneity, combined with matching performs better than a cross-sectional matching approach.

There are no design-replication studies that include inverse probability weighting (IPW) and trimming rules. Both approaches attempt to address covariate imbalance and common support problems; IPW by re-weighting the units by their propensity scores and trimming rules by estimating the optimal range of extreme propensity scores to drop from the sample (Crump et

[17] These studies have a single pre-treatment observation to form a panel.

al., 2009). Busso et al. (2009) and Busso et al. (2011) evaluate the performance of IPW and trimming with simulated data and small sample sizes (100 and 500 observations). They find that IPW performs well, but trimming performs well only when the treatment effect is homogeneous.

In sum, the literature using design-replications is small but growing. Previous studies, as a whole, do not provide conclusive evidence on the performance of non-experimental designs.

3. Study Site and Experiment

In 2007, Paul Ferraro, in partnership with Cobb County Water System employees, implemented a targeted, residential information campaign as a large-scale randomized field experiment to test the effectiveness of three messages in inducing voluntary reductions in water consumption among households in Cobb County in metropolitan Atlanta during a drought. Experimental treatments were assigned in May 2007. Ferraro and Price (forthcoming) estimate the mean treatment impact on aggregated water consumption for the four months of the summer 2007 watering season (June to September). Each treatment group comprised approximately 11,700 households and the control group comprised approximately 71,600 households.

In this study we focus on two treatments: (i) the technical information message, and (ii) the social comparison message in which a household's consumption is compared to median consumption in the county (see appendix for examples). The technical information treatment only instructed households on strategies to reduce water use. The social comparison treatment included technical information that explained how households could reduce water use, social norm-based encouragement, and a social comparison in which own consumption was compared to median county consumption. We choose the technical information treatment because it had a small estimated effect, which was insignificantly different from zero and well below the policy-relevant impact threshold identified by the Water System of a 2% reduction in water consumption. In contrast, the social comparison treatment had the largest and longest-lived (>24 months) estimated treatment effect, which was more than double the policy-relevant impact threshold (Ferraro, Miranda and Price, 2011).

We attempt to satisfy the three Heckman et al. (1997, 1998a, 1998b) criteria for forming a credible non-experimental comparison group. We form the non-random control group from households in Fulton County, which is also in metropolitan Atlanta and shares borders with Cobb County. Cobb County Water System and Fulton County Water Service Division measure the outcome variable in the same way (water meters). The two counties experienced similar water pricing policies, water sources, weather patterns, state and metro regulatory environments and other regional confounding factors during the pre- and post-experiment periods. Households did not volunteer to participate in the treatment. Some households chose to reside in a county that adopted the information campaign and others chose to reside in a county that did not adopt the information campaign. Thus selection bias only arises from the pre-treatment household decision of where to reside. Other sources of bias may come from time-varying unobservable factors specific to each county. For households in both counties, we have data on monthly pre-treatment outcomes and on a set of household and neighborhood variables that are plausibly related to both household water use and decisions to live in one county or the other. Note that the treatment is unlikely to have affected the post-treatment composition of the two counties.

Although studies comparing firm, worker or household behavior in neighboring administrative units are common in the social science and policy literatures, one might question how many independent observations one has in such contexts. In our case, do we have two observations (Cobb County and Fulton County) or thousands (the households in each county)? Following the predominant approach in the literature, we will assume that the household is the unit of observation, but will let our design-replication study inform us about the validity of this assumption.

4. Non-Experimental Designs

We use two non-experimental designs based on the outcome variable analyzed: the post-experiment, cross-sectional difference in outcomes and the difference-in-differences using pre- and post-experiment data. We describe the assumptions required for unbiased estimators of treatment effects and then describe statistical methods for making these assumptions plausible (e.g., trimming, matching, ordinary least squares regression, fixed-effects regression, inverse probability weighting). These methods offer different reweighting schemes to make the control and treatment group comparable. Previous design replication studies focus on these methods separately. By combining all them together, we provide a more systematic review of their relative performances.

The experimental design described in the previous sections permits the estimation of the Average Treatment Effect (ATE), which given treatment randomization, is equal to the Average Treatment Effect on the Treated (ATT) and the Intention to Treat (ITT). These treatment effects are defined as:

$$\mathrm{ATE} = \mathrm{E}[\mathrm{Y}^{\mathrm{T}} - \mathrm{Y}^{\mathrm{C}}] \qquad (1a)$$

$$\mathrm{ATT} = \mathrm{E}[\mathrm{Y}^{\mathrm{T}} - \mathrm{Y}^{\mathrm{C}} | \mathrm{T} = 1] \qquad (1b)$$

$$\mathrm{ITT} = \mathrm{E}[Y_z^T - Y_z^C] \qquad (1c)$$

where Y is the outcome of interest, Y^T and Y^C are potential outcomes under treatment (Y^T) and control (Y^C) conditions, T is the treatment status variable ($T=1$ indicates assignment to treatment and $T=0$ indicates assignment to the control condition) and Z represents the sample at which the treatment is administrated regardless of the treatment actually received. The household subscript i is suppressed. Given this context, using the non-experimental data, we wish to estimating the effect of the treatments on households in Cobb County; i.e., the ATT (which equals to the ITT since the program has 100% compliance).

4.1. Single Difference

One of the simplest approaches to estimate the treatment effect is taking the raw average difference between treatment and control units. This is the single difference (*D*) and can be estimated using equation (2a). The key assumption in order to provide an unbiased estimator of the ATT is that the expected outcome of the control group represents the expected outcome of the treated group in the absence of treatment (equation (2b)).

$$\mathrm{D} = \mathrm{E}[\mathrm{Y}^{\mathrm{T}} | \mathrm{T} = 1] - \mathrm{E}[\mathrm{Y}^{\mathrm{C}} | \mathrm{T} = 0] \qquad (2a)$$

$$\mathrm{E}[\mathrm{Y}^{\mathrm{C}} | \mathrm{T} = 1] - \mathrm{E}[\mathrm{Y}^{\mathrm{C}} | \mathrm{T} = 0] = 0 \qquad (2b)$$

In a non-experimental setting, equation (2b) is implausible. Thus we use four different methods that reweight the control group to control for selection on observable characteristics (or on unobservable characteristics correlated with the observable characteristics): matching, trimming, inverse probability weighting (IPW), and ordinary least square (OLS). All of these methods are different approaches, and sometimes used in combination, to replace equations (2a) and (2b) with equations (3a) and (3b):

$$D(X) = E[Y^T|X, T = 1] - E[Y^C|X, T = 0] \qquad (3a)$$

$$\mathrm{E}[\mathrm{Y^C}|\mathrm{X}, \mathrm{T} = 1] - \mathrm{E}[\mathrm{Y^C}|\mathrm{X}, \mathrm{T} = 0] = 0 \qquad (3b)$$

where *X* are observable covariates. In other words, conditional on the variables in *X*, the expected outcome of the control group represents the expected outcome of the treated group in the absence of treatment and thus *D(X)* provides an unbiased estimator of the ATT. The zero selection bias assumption is equivalent to assuming that unobservable variables affecting potential outcomes in the control condition are uncorrelated with exposure to treatment, also known as a selection on observables assumption (Ravallion, 2008).

Matching methods match treated units to non-experimental control units in order to achieve balance on the means and distributions of confounding covariates, thereby re-establishing experimental conditions where the only systematic difference between the groups conditional on the covariates is exposure to the treatment (Blundell and Costa-Dias, 2009). We select the matching method that achieves the best covariate balance (see section 8). We match with and without calipers. Calipers can further improve covariate balance between treatment and control groups by defining a tolerance level for judging the quality of the matches; if a treated household does not have a match within the caliper (i.e., available controls are not good matches), it is eliminated from the sample. Calipers reduce bias, but at the cost of estimating a treatment effect on a subsample that may not be representative of the population of treated households. Sekhon (2010) recommends using calipers instead of simply enforcing a common support because calipers drops outliers and inliers while common support only drops outliers. Rosenbaum and Rubin (1985), in the context of propensity score, found that Mahalanobis matching with calipers is superior to Mahalanobis matching alone in reducing bias. As a final measure to control for selection on observables, we use a post-matching bias-correction procedure that asymptotically removes the conditional bias in finite samples (Abadie and Imbens, 2006).

Ho et al. (2007) argue that matching methods can also be viewed as a way to pre-process data to make identifying assumptions more plausible and make treatment effect estimates less dependent on the specific post-matching statistical model. Since the adjustment for covariates is done non-parametrically with matching, the potential for bias is greatly reduced compared to parametric analyses (Ho et al., 2007). Imbens and Wooldridge (2009) also recommend combinations of methods to achieve robustness under misspecification, a common problem in parametric models. We thus also use matching as a way to pre-process the data before applying a parametric estimator (e.g., ordinary least squares, fixed effects). Given we are interested on estimating the ATT, there can be gains to dropping control units that do not represent accurate counterfactuals for treated units. Rosenbaum and Rubin (1985) suggest that including all control units (which in our case are many more than treated units) can artificially deflate the standard errors.

Like matching, trimming is an approach that reweights the control group by restricting the sample. The trimmed sample discards (or applies a weight of zero) to observations with extreme propensity scores. Trimming is commonly applied when there is limited overlap on covariates. Limited overlap introduces substantial bias and large variance (Crump et al., 2009). Although previous studies drop observations with extreme propensity score values in an ad-hoc manner, Crump et al. (2009) derive an optimal trimming rule for discarding observations with extreme propensity scores. The subset of observations is defined in terms of the distribution of covariates and the treatment indicator without depending upon the distribution of the outcome variable. This practice, however, estimates a treatment effect within a subpopulation that may not be representative of the whole population.

While covariate and propensity score matching can eliminate bias, inverse probability weighting (IPW), which uses the propensity scores as weights, can be more efficient asymptotically (Hahn 1998; Hirano and Imbens, 2001; Hirano et al., 2003; Busso et al., 2009). IPW corrects for unrepresentative, non-random sampling of potential outcomes by giving less weight to those individuals who have a high probability of treatment, conditional on the set of observable covariates. The IPW estimator weights observations by the inverse of nonparametric estimates of the propensity score ($\hat{p}(X_i)$). Equation (4) and equation (5) shows the weights for the average treatment effect (ATE) and average effect on the treated (ATT), respectively.

$$Weights_{ATE} \begin{cases} 1/\hat{p}(X_i) & if\ treated\ unit \\ 1/1-\hat{p}(X_i) & if\ control\ unit \end{cases} \qquad (4)$$

$$Weights_{ATT} \begin{cases} 1 & if\ treated\ unit \\ \hat{p}(X_i)/1-\hat{p}(X_i) & if\ control\ unit \end{cases} \qquad (5)$$

We focus on the ATT weights because we are interested in estimating the ATT. However, IPW studies in the literature have focused on the ATE weights, e.g., Emsley et al. (2008), Hirano and Imbens (2001), Millimet and Tchernis (2009). Thus we also present results using the ATE weights.

Hirano et al. (2003) show that IPW yields a fully efficient estimator for the ATE. However, IPW is sensitive to large values of the estimated propensity score. Suppose there are observations with specific covariates that result in a high propensity score (e.g., 0.9). This suggests that for any 10 observations with the same covariate value, 9 of them will be treated and 1 of them will be untreated. To be representative of the treated group, the untreated observation with that propensity score must be counted 9 times. Thus, the overall treatment effect estimated may end up being very sensitive to observations with large weights (Hirano and Imbens, 2001; Huber et al., 2010). Some authors (e.g., Stuart, 2009) have thus recommended combining trimming with IPW (in other words, applying IPW to a trimmed sample should provide more unbiased estimator than applying it to the full sample).

Using covariates *X*, we can also estimate equation (3a) using ordinary least squares (OLS). OLS is a parsimonious method and provides a reasonable approximation of the treatment effect. OLS provides a conditional-variance-weighted estimator of ATE, which is neither equal to ATE nor ATT. Angrist and Krueger (1999) suggest “unbiasedness on least square is attainable when the variables that determined the treatment assignment are known, quantified, and included in the equation”. Some studies have shown OLS regressions perform relatively well in replicating experimental benchmarks (Shadish et al., 2008; Arceneaux et al., 2006). Unlike matching and trimming, however, OLS fails to alert the analyst to a lack of common support between treated and control units. Moreover, OLS imposes a particular functional form that, if different from the true form, may introduce additional sources of misspecification.

4.2. Difference-in-Differences

The second non-experimental design is the difference-in-differences (*DD*) design. *DD* consists of taking the average difference between treatment and control units for periods before and after the program were implemented (e.g., McKenzie et al., 2010). *DD* only requires two points in time (before and after). If more time periods are available, a panel data estimator can be used. Under certain assumptions, *DD* removes bias produced by time-invariant unobservable differences between treatment and control groups and it cancels out upon subtraction (Blundell and Costa Dias, 2009; Ravallion, 2008). When using the raw, unadjusted difference-in-differences, we refer to this estimator as the Unconditional Difference-in-Differences (UDD), represented as equation (6a), where *0* represents the pre-treatment period and *1* represents the post-treatment period. UDD does not impose any particular functional form. The key underlying assumption under this approach is that the outcome trend in the control group represents the counterfactual trend of the treated group in the absence of treatment (equation (6b)).

$$DD = E[Y_1^T - Y_0^T | T = 1] - E[Y_1^C - Y_0^C | T = 0] \qquad (6a)$$

$$E[Y_0^T - Y_0^C | T = 1] = E[Y_0^T - Y_0^C | T = 0] \qquad (6b)$$

In other words, in the absence of the treatment the expected potential changes in both groups are equal. Greenstone and Gayer (2009) suggest that this strong assumption is likely to be invalid in settings where behavioral responses are possible. We examine this assumption more thoroughly below.

Similar to the single difference estimator, difference-in-differences can be estimated conditional on observable covariates using equation (7a), thereby weakening the identifying assumption to (7b).

$$DD(X) = E[Y_1^T - Y_0^T | X, T = 1] - E[Y_1^C - Y_0^C | X, T = 0] \quad (7a)$$

$$E[Y_0^T - Y_0^C | X, T = 1] = E[Y_0^T - Y_0^C | X, T = 0] \qquad (7b)$$

Difference-in-differences is a particular case of panel data. With more observations before and after treatment assignment, one can use a panel data estimator version of the Equation (7a). Given the monthly measurements of water use in our data, we believe a panel data estimator would be the most appropriate non-experimental estimator. Such estimators can control for time-invariant unobservable household characteristics that are associated with treatment assignment and water consumption. Among panel data, the standard fixed-effects estimator (FEPD) is the workhorse in empirical studies on linear panel data models (Wooldridge, 2005). The main idea behind fixed-effects identification strategies is to use repeated observations on households to control for unobserved and unchanging characteristics that are related to water consumption and the treatment variable (Angrist and Krueger, 1999).

The key identifying assumption for panel data estimator is that the water consumption trends would be the same in both counties in the absence of treatment, the same assumption as for difference-in-difference approach.[18] If the outcome variable pre-treatment trend for the control and treatment group look similar, then it is plausible to assume that unobserved characteristics affecting program participation do not vary over time with treatment status. However, the implicit assumption that the unobserved effect is additive and time invariant is, as Angrist and Krueger (1999) pointed out, "arbitrary in the sense that it usually does not come from economic theory or from information about the relevant institutions".

There is an additional but very important assumption on fixed effect panel data models to consistently estimate the treatment effect: the treatment effects are constant (homogenous) and additive (Gibbons et al., 2011; Imai and Kim, 2011). Fixed effects regressions weights on sample variances rather than sample frequencies. More specifically, fixed effects regressions average the group-specific coefficients proportional to both the conditional variance of treatment and the proportion of the sample in each group. Thus, if the assumption of a homogenous treatment effect does not hold, then the fixed effect estimator overweights groups that have larger variance of treatment conditional upon other covariates and underweights groups with smaller conditional variances (Gibbons et al., 2011). For a clear exposition refer to Gibbons et al. (2011).

[18] In the difference-in-difference approach, where there is only one observation pre-treatment, this assumption is difficult to test.

4.3. Summary of Non-experimental Designs and Methods

Table 9 shows all the combination of non-experimental designs and methods used in this study. The full sample does not reweight the control group observations. The trimmed sample discards (or applies a weight of zero) to observations with extreme propensity scores. The matching sample (without and with calipers) restricts the sample to control households that are more similar to the treated households. For these samples, we estimate raw differences, matching, inverse probability weighting, ordinary least squares regression, and fixed-effect panel data regression.

As we stated before, applying IPW to a trimmed sample should provide more unbiased estimator than applying it to the full sample. In the same vain, we expect that OLS will perform better when combined with the matched and the trimmed samples because (a) matched and trimming will discard inappropriate counterfactuals for the treated group, and (b) OLS will help to control for other imbalances that cannot be handled with the matching and trimming method alone.

Regarding fixed effect panel models, we expect that its performance can be evaluated from two viewpoints. First, if the pre-program trends do not look similar, then we expect that the non-experimental method would do a poor job on replicating the experimental estimates. However when samples are more similar (e.g., under the matched sample or the trimmed sample) *and* that improvement is reflected on the pre-program trends, then we expect that panel data will improve its performance. Second, Ferraro and Miranda (2011) found that indeed there is evidence of heterogeneous responses to the social comparison treatment. Thus, after matching, the performance of the fixed-effects estimator should improve, but under caliper matching, assuming the resulting sample is not unrepresentative of the original sample, it should perform best because the variance between treatment and control units is reduced.

5. Data

The data for our analysis are derived from four sources. Monthly water consumption at the household level is obtained from Cobb County Water System (experimental data) and from Fulton County Water Service Division (non-random comparison group). We have thirteen months of pre-experiment data (May 2006 to May 2007) and four months of post-experiment data (June to September 2007).[19] Covariates variables are obtained from county tax assessor databases and the 2000 US Census. Tax assessor databases provide home and property characteristics. The US Census provides data on neighborhood characteristics at the block group.[20]

We select covariates that are observable to policymakers and that theory or empirical studies suggest could be important controls in a study on water conservation. Ferraro and Price (forthcoming) analysis shows that previous water use predicts future water use. Based on metering data from the water utilities, we use the two variables used by Ferraro and Price in their analysis: June – November 2006 billed use (corresponds to May – October use, which is the main water use season) and April – May 2007 billed use (to reflect changes in landscaping prior to treatment assignment in May 2007).[21]

From the 2007 county tax assessor databases we select fair market value of home ($), property size (acres), and age of home (years). We include fair market value (as a proxy for income and wealth) because Ferraro and Miranda (2011) found that high-income households are more responsive to non-pecuniary, information-based messages. Two measures that reflect the scope and incentives for water conservation are the age of the home and the size of property. Older homes, on average, have older water-intensive capital, which are more cost-effective to replace to achieve water conservation goals. On the other hand, bigger properties houses are strongly correlated with the level of water use.

From the 2000 US Census, we choose block group measures of per-capita income ($), percent of adults over 25 years old with college education or higher, percent of people living below poverty line, percent of population that is white, and percent of renter-occupied housing units. Environmental preferences of household occupants would likely also affect their treatment responses. However, we (and water utilities) cannot observe environmental preferences, but survey evidence suggests that environmental preferences vary with education levels and race.

[19] Administrative data correspond to water billing month. Thus, for example, our pre-experiment data are for June 2006 to June 2007 billing periods.

[20] A block group is a subdivision of a census tract. The number of people in a block group varies from 600 to 3,000 people, with an average size of 1,500 people.

[21] Later, in an Appendix we reevaluate the use of previous water use by considering only summer 2006 (June to September use).

Thus, we use measures of education (percent with bachelor's degree or higher) and race (percent white) at the census block group. We also include a variable regarding renters/owners. Ferraro and Miranda (2011) show that owner-occupants have greater social connections to their neighbors and thus be more responsive to pro-social messages. Finally, we include variables regarding poverty measure in order to capture differences across counties (in the next paragraph we show strong differences between Fulton and Cobb County on poverty measures: Fulton doubles on poverty to Cobb County).

Table 3 shows descriptive statistics for Cobb County and Fulton County based on the 2000 US Census. Both counties have similar proportion of urban population (99% and 98%, respectively), average household size (2.6 and 2.4 respectively) and population over 25 years with higher degree of college (68% and 65%, respectively). However the Cobb County population comprises more White citizens (72% versus 48%) and its income distribution is quite different. Although Cobb County has lower per capita income ($ 27,863 versus $ 30,003), it also has less poverty (6% versus 16%) and a higher median income ($58,289 versus $47,321).

The outcome variable is water use. We have data on 66,849 residential households in Fulton County. Cobb County measures water consumption monthly. However, Fulton County measures it bimonthly or, for some households, less frequently. We split Fulton County billed use across months (e.g., if June bill was 2000 gallons, 1000 was assigned to May and 1000 to June; more sophisticated weighting that takes into account seasonal variation in use did not affect the estimates and thus is not reported). We remove from the Fulton County sample any household that went for 5 or more months without a bill during the study period (2,888 obs.). To make the Fulton sample comparable to the eligible set of households for randomization in the Cobb experiment, we also remove from the Fulton sample any households that consumed fewer than 20,000 gallons between May and September 2006 (14,865 obs.), any households that saw a change in the name of the billed customer between May 2006 and April 2007 (1,234 obs.), any households that consumed more than 1,000,000 gallons between May and October 2006 (a sign of catastrophic leaks; 7 obs.), and any households not actually within Fulton County boundaries (376 obs.). See Ferraro and Price for explanation of the experimental sampling design.

Table 11 shows average water consumption in thousands of gallons during key watering seasons for Cobb households in the experiment (including the randomized control group) and

Fulton County households (the non-experimental control group). There are differences across the counties, but within the experimental groups in Cobb County, there are no significant differences. Regarding pre-experiment data (summer 2006, winter 2006-2007, and spring 2007) in Cobb County, there are no statistical differences across treatments and random control group. The only difference arises in post-experiment (summer 2007) for the social comparison treatment. The unconditional average treatment effect is approximately 4.2% reduction in water consumption, which is statistically significant at p=0.01, while the unconditional reduction for the technical information treatment is 0.02%, not statistically significant at p=0.10.

Finally, Table 14 shows descriptive statistics from the tax assessor and the US census for our specific sample. In general, Cobb County has slightly cheaper, older and smaller properties. Further, Cobb County has lower per-capita income, lower percentage of people with higher degree, lower percentage of renter occupied housing units, and a lower percentage of white.

6. Experimental Benchmark

As is implicitly assumed in previous design-replication studies, we assume the experimental estimate is the relevant benchmark, despite the presence of sampling error. Given the characteristics of the available data and the context in which the program took place, we can take advantage of the cross-sectional and panel data structures.

With the cross-sectional data, we estimate an OLS regression using equation (8). To increase statistical precision and to properly estimate treatment effects given the within-meter route randomization design, Equation (8) includes observable covariates X_i described previously (previous water use, household and neighborhood characteristics) and dummy variables for the meter routes:

$$Water\ Use\ Summer\ 2007_i = \alpha_i + \beta * Treatment_i + \delta' X_i + \varepsilon_i \qquad (8)$$

With the panel data, we estimate equation (9). The dependent variable is monthly water consumption and the time variable goes from May 2006 to September 2007. Post-experiment months are the 2007 summer months (June 2007 – September 2007). Estimating treatment effects in the experimental design does not require using equation (9) (in fact it needlessly adds assumptions), but we believe the most appropriate non-experimental design would be one that takes advantage of the panel data structure to control for time-invariant, unobservable household characteristics that are associated with treatment assignment and water consumption. Thus when comparing estimates from panel data estimators to the experimental benchmark estimate, we wish to compare "apples to apples." Potential approaches to estimating (9) include first differencing and fixed effects estimation (the key identifying assumption of the random effects estimator does not seem tenable in this context: i.e., unobserved, household time-invariant effects that are uncorrelated with the relevant explanatory variables in all time periods). The fixed effects are assumed to exist at the level of the household.

$$Monthly\ Water\ Use_{it} = \alpha_{it} + \beta * Treatment_{it} + \mu_i + \eta_t + \varepsilon_{it} \qquad (9)$$

The experimental benchmark estimates using equation (8) are shown in Table 13. The analysis in Table 13 essentially replicates Ferraro and Price (forthcoming) using the sample for which we could match addresses in the utility, tax assessor and census databases, and adds additional controls that are used in the other analyses.[22] The results shown in Column (2) and (3) are nearly the same as in Ferraro and Price (forthcoming): the social comparison message reduces summer 2007 water consumption, on average, by about 1700 gallons ($p<0.01$) and the technical information message reduced consumption, on average, by only about 130 gallons ($p>0.10$).

Table 14 presents the experimental benchmark estimates using equation (9). The important variables are the interaction term between the treatment and the post-experiment period. Results show that the technical information message reduced monthly consumption, on average, by very little, whereas the social comparison message reduced monthly consumption, on average, by about 350 gallons, which are very close to the estimates reported in Ferraro and Price when multiplied by four. Angrist and Krueger (1999), Angrist and Pischke (2009), and Swaffield (2001) suggest that panel data estimators are smaller than estimates from cross-sectional data due to bias from measurement error that is aggravated when individual effects are removed.

[22] 13.8% of experimental data were dropped. Unconditional differences on pre-experiment water consumption are statistically different between observations kept in the analysis and observations dropped for each treatment and the experimental control group. Average values for dropped observations are higher than kept observations, however when controlling for other observable covariates experimental results do not vary.

7. Non-experimental Evaluation Criteria

The aim of our study is to evaluate the performance of non-experimental (or observational) empirical designs in the context of the Cobb County conservation information campaign. The design-replication literature provides no guidance for determining whether a non-experimental estimate is "good" based on the experimental benchmark estimate. Many studies apply no criteria at all. We define two criteria for judging the quality of the non-experimental estimate:

> *Criterion 1*: The non-experimental point estimate should be in the 95% confidence interval of the experimental point estimate. This criterion is applied in previous design-replication studies (e.g., Arceneaux et al., 2006; Greenberg et al., 2006). We also wish to make the correct inference in testing the null hypothesis of no treatment effect (Type 1 error = 5%).
>
> Based on the cross-sectional results (Table 13, column 3), *Criterion 1* implies that the non-experimental estimate for the social comparison treatment should be on the range [-2081, -1383] and, for the technical information treatment, on the range [-524, +276]. Based on the panel data results (Table 14), *Criterion 1* implies that the non-experimental estimate for the social comparison treatment should be on the range [-443, -263], and for the technical information treatment, on the range [-111, +99].[23]
>
> *Criterion 2*: To be considered the "best" non-experimental estimator, an estimator must satisfy two properties: (a) under random sampling using bootstrapping, it has the highest concentration of estimates within the 95% confidence interval of the experimental estimate; and (b) the concentration of estimates with the 95% confidence interval of the

[23] Another criterion could be defined as a fraction of the policy-relevant treatment effect. One would not want to encourage policymakers to scale-up a treatment that in fact does not have a policy-relevant treatment effect, fail to encourage scale-up of a treatment that has a policy-relevant effect, or grossly overestimate how effective any scaled-up initiative will be (e.g., we would not want to claim a treatment effect 25% greater than the experimental benchmark). In our case, the minimum relevant treatment effect defined by the water utility is a 2% decline in average summer water use and thus if an estimate satisfies Criterion 1, we would not encourage scaling-up of the technical information treatment and would encourage the scaling up of the social comparison treatment. Given this fact and that we have no policy guidance on an upper bound threshold (i.e., what comprises a "gross overestimate"), we do not attempt to craft a third criterion.

experimental estimate should be similar for both treatments (i.e., an estimator cannot be judged to be successful if it can only consistently recover one of the treatment effects). Bootstrapping can be used to estimate a distribution of treatment effects, thereby allowing one to judge the sensitivity of the proposed methods to different samples (McKenzie et al., 2010). Sensitivity to sample choice has been a concern in the design-replication literature (e.g., Smith and Todd (2005); Dehejia and Wahba (1999, 2002) debate).

8. Non-Experimental Sample

Recall we use three samples in our non-experimental analyses: the "full sample" using all the observations, a "trimmed sample," and a "matched sample." The trimmed and matched samples were created using the full set of covariates described in Section 5. To create the trimmed sample, we use a logistic regression and the full set of covariates to calculate the optimal trimming rule (Crump et al. 2009). The rule implies observations with estimated propensity scores outside the interval [0.06, 0.94] should be discarded (i.e., assigned a weight of zero).[24] Figure 4 shows clearly the observations discarded.

To create the matched sample, we choose the matching method that generates the best covariate balancing results for our sample: Mahalanobis covariate matching. We apply this matching method with and without calipers of 1 standard deviation. Figure 5 shows the observations discarded based on covariate matching with calipers.

Based on the arguments in Section 4, we expect that the 'matched' sample will perform better than the 'trimmed' sample, and the latter better than the 'full' sample, because of improvements in covariate balance among treated and control units. The full sample includes both outliers and inliers with respect to the treated group. The trimmed sample drops only

[24] The rule for both treatments is similar, but not identical. The exact optimal number for Social Comparison Treatment is 0.06116, while the optimal number for the Technical Information Treatment is 0.06227. Logit and Probit estimates differ by 1.6 units. Logit coefficient is 1.6 times larger than Probit coefficient (Train, 2009). Because we are interested to understand which method provides closer results to the experimental one, the selection of Probit and/or Logit matters. It will influence through the propensity score for the trimming set, but also it will influence the weights used (either ATT or ATE). Thus later, in an Appendix, we provide some evidence about the effect when choosing one of the two models.

outliers. The matched sample drops outliers and inliers (bad counterfactuals). Trimming and matching, however, lead to estimates of treatment effects for subpopulations rather than the entire relevant population.

Before applying our empirical methods in the next section, we compare balance on the covariates selected and evaluate whether there is improvement across samples. Table 15 and Table 16 show these balancing results for the social comparison and technical information treatment, respectively. Column (1) shows results for the full sample, column (2) for the trimmed sample, and column (3) and (4) for the matched samples. Column (5) to (7) shows the percentage improvement from the balance achieved in the full sample.

For each covariate we show four ways to evaluate the improvement in covariate balance: (a) difference in means; (b) standardized mean difference, for which Rosenbaum and Rubin (1985) suggest that a standardized difference greater than 20 should be considered large (Lee, 2011); (c) eQQ mean difference, a non-parametric test that evaluates the rank rather than the precise value of the observations (Ho et al., 2007); and (d) variance ratio between treated and control units, which should be equal to one if there is perfect balance (Sekhon, 2011).

Out of ten covariates, five of them show clear improvement across all four measures as one moves from the full sample to the trimmed sample, matched sample, and caliper matching sample (aggregate water use from May 2006 to October 2006, fair market value of home, percent of adults over 25 years old with college education or higher, per-capita income, percent of renter-occupied housing units). Among these five covariates, three of them reduced their standardized mean difference from above 20 to less than 20, in absolute value. Three other covariates (age of home, percent of people living below poverty line, percent of population that is white) see improvements moving from the full sample to the trimmed sample, but declines with matching (without caliper), based on standardized mean difference. Caliper matching, however, improves balance on all four measures. These three covariates reduced their standardized mean difference above 20 (from the full sample) to less than 20 in absolute value (caliper matching sample), and the variance ratio is closer to one for the caliper-matching sample. The final two covariates (aggregate water use for March and April 2007, property size) see a mix of improvements and declines, but the original values were not substantially unbalanced. The balancing results thus corroborate our expectations trimming and matching, in general, improve covariate balance and caliper-matching shows the best balance.

9. Non-Experimental Results

In this section, we present the results for each design and method combination described in Table 9. Table 17 and Table 18 show results for the single difference and the difference-in-differences for the social comparison and technical information treatments, respectively. Column (1) and Column (3) show the treated and control average consumption after the experiment (Summer 2007), while Column (2) and Column (4) show the treated and control average consumption before the experiment was implemented (Summer 2006). Based on these estimates, Column (5) calculates the difference-in-differences for treatment and control units before and after the experiment took place. Similarly, Column (6) shows the single average difference post-experiment between treated and control units. Furthermore, the first panel of results shows raw mean differences, the second panel of results shows matching with bias correction results, the third and fourth panel of results shows weighted means using IPW for ATT and ATE, respectively.

Remember that the average treatment effect in the experimental design implies a reduction of about 1.7 thousands of gallons for the social comparison treatment, and about 0.12 thousands of gallons for the technical information message (see Table 13). All non-experimental estimates for unconditional difference-in-differences overestimate the experimental benchmarks and, more importantly, none of them imply a reduction in water consumption (see Column (5) in Table 16 and Table 17). These results are striking. It suggests that there are other relevant differences across counties, besides the treatment effect, that we are not conditioning for. Thus the difference-in-differences design does not provide a good estimate of the experimental benchmark. Major discrepancies come from the difference in summer 2006 between treated and control groups, implying that the key identifying assumption (i.e. that the mean trend of the control group represents the proper counterfactual for the treated group) does not hold in our case. The non-experimental estimates provide misleading information: they suggest that both treatments increase, rather than decrease, water consumption. Naturally, these estimates do not satisfy *Criterion 1*.

Estimates based on single difference estimators for Summer 2007 are shown in Column (6) in Table 16 and Table 17. The results are less biased than difference-in-differences estimates. Simple difference shows the right sign for the social comparison treatment (Table 17): a reduction in water consumption, rather than an increment. Under matching sample (no calipers) and using the full set of covariates, the single difference for both treatments provides closer results to the experimental benchmark than the full, trimmed and caliper-matching samples.[25] IPW does not improve weighted differences using ATT or ATE weights. In fact, the estimate is worse when combining matched samples with IPW method: the results are in opposite direction to the experimental effects. Thus, these results suggest that combining matching to pre-process the data and then re-weighting the matched sample with inverse propensity score weights is not an appropriate mix of methods.

Table 19 and Table 20 show single difference OLS results for social comparison and technical information treatment, respectively. Column (1) and (2) show results for the full sample. Columns (3) and (4) show results for the trimmed sample. Column (5) to (8) show results for the matched sample, both without and with calipers. Results show that specification matters substantially. The estimates and inferences drawn are highly sensitive to the set of covariates included.[26] Just including previous water use as covariates OLS in all samples overestimate the experimental effects for both social comparison and technical information treatment. This suggests that there are other important covariates not including in the regression that affect water consumption. It is clear that using baseline water use alone provides inadequate control under selection on observables.

The census and tax assessor variables are critical to obtaining estimates closer to the experimental estimates. When including the full set of covariates, results improve substantially and predict that there is a reduction of about 1,550 gallons for the social comparison treatment

[25] However, these results are very sensitive to covariates specification. See Appendix 10 for more matching results using different set of covariates. It is worth nothing that only using neighborhood variables combined with caliper matching and bias correction, results are similar to the experimental benchmark. However, again, these results are very sensitive to matching specifications; moreover when literature suggest that pre-treatment variables are important in the matching context to recover experimental estimates.

[26] Table 11 and 12 show two sets of covariates: (a) only previous water, (b) previous water use, property, and neighborhood variables. Appendix 11 and Appendix 12 show other sets of covariates: (c) only property variables, (d) only neighborhood variables, and (e) property and neighborhood variables.

(the experimental estimate shows a reduction of 1,700 gallons) and no statistically significant treatment effect for the technical information treatment (similar to the experimental estimate). These results are comparable using the full sample, the matching sample and the caliper-matching sample: estimates predict a statistically significant reduction for the social comparison treatment and statistically insignificant effect for the technical information treatment.[27] Thus, using OLS regression with single difference and the full set of covariates, the full sample and the matched sample (with and without calipers) satisfy *Criterion 1*.

OLS results using the trimmed sample, however, are striking and difficult to interpret. They differ substantially from the estimates using the other three samples. In both treatments, besides the improvement on balance among covariates, they show an increment on water consumption, rather than reduction. Further, under some specifications, we find substantial biases and statistically significant. Remember that for this exercise, observations with estimated propensity score outside the interval [0.06, 0.94] are discarded. We are not yet applying inverse propensity score weights. Adding them might provide insights into why OLS and trimming does not provide results closer to the experimental benchmark.

Before showing the IPW results, Table 21 and Table 22 illustrate difference-in-differences OLS results for social comparison and technical information treatment. In none of the estimations under different samples are results close to the experimental benchmark. As with the unconditional difference-in-differences results, the identification assumption does not appear to hold in our case and conditioning on other relevant covariates does not help to make this assumption more plausible.[28]

Following the cross-section estimates, we examine more deeply the performance of IPW and the type of weights applied. IPW is subject to extreme values, thus we only show results for the

[27] As we mention before, these results vary greatly. Appendix 11 and Appendix 12 shows that using only property and neighborhood variables, the treatment effects are positive and significant for both treatments.

[28] Table 13 and Table 14 were estimated also without using covariate "Water Use from June - November 2006" because it could be argued that this period of water use includes Summer 2006 that is also in the left-hand side of the regression. However, results (not shown) are more biased than the presented in these two tables.

trimmed sample.[29] Figure 6 shows that the estimated propensity scores for each treatment are condensed in the extreme of the distribution (closer to zero and one).

Before using IPW, we compare the balance on the set of covariates using ATT and ATE weights. Results are shown in Table 23 and Table 24 for the social comparison and technical information treatment, respectively. These results are based on the trimmed sample. Column (1) and (2) show raw means for treatment and control units. Column (3) and (4) show weighted means using ATT weights. Column (5) and (6) show weighted means using ATE weights. Columns (7) – (9) show the difference between control and treated groups for each type of weights (no weights, ATT weights and ATE weights, respectively).

Despite our preference for using ATT weights, ATE weights achieve systematically better balance between the treated and control groups for both treatments (see columns (7) to (9)). ATE weights reduce imbalance for all 10 covariates. ATT weights only reduce imbalance on 7 covariates (imbalance increases for water use in April and May 2007, age of home, percentage of renters). Notice that the mean average for the treated observations in column (1) and (3) are equal. This is because the ATT weights for the treated group is equal to one and this could explain the better balance using ATE weights: treated units are also weighted to be more similar to control units. Therefore, better balance can be achieved using ATE weights.

Table 25 shows non-experimental estimates using IPW. Both types of weights overestimate the effect of the social comparison (ATT estimate: -2.9; ATE estimate: -2.3) and technical information treatment (ATT estimate: 0.1; ATE estimate: 0.8). The difference between ATT and ATE estimate for each treatment are not statistically different from each other, but ATE weights are closer to the social comparison experimental benchmark and ATT weights are closer to the technical information experimental benchmark. Thus, IPW method improves the performance on the trimmed sample (without weights), but still does not satisfy *Criterion 1*.[30]

[29] Results using full sample show that coefficients are subject of substantial biases. For example, when using ATT weights the estimated coefficient for the social comparison treatment is -25.7, and the coefficient for the technical information treatment is -18.2. These results corroborate the initial idea that IPW is very sensitive to observations with large weights, thus using the trimmed sample is the best approximation when using IPW.

[30] Results shown are based on Logit models. Logit coefficients are roughly 1.6 times larger than Probit coefficient (Train, 2009). Appendix 13 shows results using Probit models instead of Logit models. Despite smaller Probit coefficients, there is no linear relationship in the estimates on Appendix 13 when compared to Table 17. This

Finally, we evaluate the performance of non-experimental techniques using the panel data structure. As we mentioned earlier, we believe that panel data is the right empirical approach given the nature of the data. Under homogenous treatment effects, panel data under the full sample should provide results close to the experimental benchmark. Under heterogeneous treatment effects, panel data in the caliper-matching sample should perform best.

Prior to presenting the estimates, we examine the assumption of parallel-trends prior to the experiment for treated and control units. Figure 7 shows the mean monthly consumption trends for the social comparison treatment, technical information treatment, Cobb County random control (given randomization, their trends look identical) and the Fulton County non-random control group. The two lines do not look similar, thus suggesting that the Fulton households, on average, do not form a good counterfactual for the treated Cobb households. Thus, we do not expect to have results close to the experimental benchmark using the full sample. Figure 8 and Figure 9, for the social comparison and technical information, respectively, shows that the trend lines become much more similar after matching and that caliper-matching does not affect the mean pre-treatment trends by much. As we will see, however, the use of calipers does have a substantial impact on the treatment effect estimates. This impact suggests that similar mean trends before the experiment is a necessary, but not sufficient, condition to ensure the assumptions of the fixed panel data estimator is satisfied.

Table 26 shows results for the full sample, trimmed sample, matching, and caliper-matching samples. Using the full sample, we estimate that the social comparison message reduced average use by 965 gallons per month and the technical information message reduced average use by 618 gallons per month. The non-experimental estimate of the social comparison treatment effect is almost three times larger than the experimental estimate, while the non-experimental experimental estimate of the technical information treatment effect is over ten times larger. These estimates do not satisfy *Criterion 1*. Importantly, they send the wrong signal to decision makers: they erroneously imply that the technical information message achieves almost two-thirds of the average impact of the social comparison message. Given technical information

happens because the choice model influence in two ways: (a) through the optimal selection rule and trimming set, and (b) through the weights used. Having said that, Appendix 13 shows that ATT weights increase bias to the non-experimental estimates, while ATE weights reduce bias to the non-experimental estimates. Later, on the bootstrapping section we reevaluate the difference between ATT and ATE weights for IPW method.

messages are (a) cheaper because they require fewer sheets of paper and fewer data (no need to examine past consumption patterns), (b) can be targeted to the entire population rather than only to customers who lived in their home the previous year, and (c) are well understood by utility managers, the results could be interpreted as implying the technical information message would be preferable.

Columns (2) and (6) present estimates using the trimmed sample. The estimates are smaller than estimates with the full sample, but still they provide the wrong information: both non-experimental estimates do not satisfy *Criterion 1*, although the technical information treatment is very close to being within the 95% confidence interval of the experimental estimate. Columns (3) and (7) present estimates using the sample pre-processed by matching. The estimates improve than the trimmed sample and are very close to satisfying *Criterion 1* (in fact, the estimate for the technical information treatment does satisfy it).

Columns (4) and (8) present estimates using the sample pre-processed by caliper matching. Despite the barely perceptible impact of caliper matching on the trend of Fulton County pre-treatment average monthly water use (see Figure 8 and Figure 9), caliper matching prior to estimation substantially improves the estimates from the fixed effect model. The estimates from the non-experimental data are nearly identical to those generated by the experimental data (see Table 12) and satisfy *Criterion 1*. Thus the non-experimental estimator does a good job of replicating the experimental estimator. Importantly, the inferences drawn from a test of the null hypothesis of zero impact are the same from experimental and non-experimental estimators.[31]

[31] An open question is regarding what pre-treatment covariates should be included in the analysis. Dehejia and Wahba (1999, 2002) included two-year pre-treatment data, thus, dropping observations that do not satisfy their requirement. Under that subsample, they found that propensity score matching did recover the experimental estimates. However, Smith and Todd (2005) suggested that their conclusions are not generalizable and depends on their choice of a particular subsample. Thus, we reevaluate our point estimates using water consumption in Summer 2006 (from July to October 2006) instead of (a) water use from June to November 2006 and (b) water use in April and May 2007. In our case does not mean a different sample, only a change in the pre-treatment covariate. Results are shown in Appendix 14. In general, we observed more biased non-experimental results using only one point in time as water pre-treatment consumption. For instance results on Section D), under OLS, we cannot recover the experimental estimation for any sample specification (moreover, results provide the wrong signal to decision makers: treatments increase water consumption). Comparing these results to Table 19 (social comparison treatment) and Table 20 (technical information treatment), we did recover the experimental estimates using the full set of covariates (pre-treatment water use, household characteristics and neighborhood variables) and for the full and matching samples. Further, another important point to make is that under fixed-effects panel data, using water use on Summer 2006 to select the matching and caliper-matching samples, helps to recover the experimental estimates

10. Non-Experimental Bootstrapping

The results in the previous section compare one experimental estimate with one non-experimental estimate. However, this comparison does not reveal anything about the distribution of the estimators (Huber et al., 2010). In this section, we apply bootstrapping methods in order to estimate the distributions of treatment effects using the non-experimental data. These distributions reveal how sensitive each evaluation design is to random sampling with the same number of observations. The bootstrapping analysis used 500 repetitions for each non-experimental method.

Table 27 summarizes the results.[32] The table reports the percentage of repetitions that satisfy *Criterion 1*. Column (1) and Column (2) show the results for the social comparison treatment and the technical information treatment, respectively. In other words, the table reports the percentage of point estimates that is within the 95% confidence interval of the experimental estimate, which relates to our *Criterion 2*. Recall that *Criterion 2* requires the "best" estimator to satisfy two properties: (a) under random sampling using bootstrapping, it has the highest concentration of estimates within the 95% confidence interval of the experimental estimate; and (b) the concentration of estimates with the 95% confidence interval of the experimental estimate should be similar for both treatments (i.e., an estimator cannot be judged to be successful if it can only consistently recover one of the treatment effects).

Under the full sample, only 22% of the estimates from the single difference estimator using OLS are within the 95% confidence interval of the experimental benchmark. The Fixed Effect Panel Data (FEPD) estimator does worse. It consistently overestimates the treatment effect for both treatments: social comparison's treatment effect estimates vary from [-1.20; -0.72], while the technical information estimates vary from [-0.92; -0.31] (results not shown).

(see Section I) on Appendix 14). Non-experimental matching results (without calipers) are closer to the 95% confidence interval experimental estimate (in fact, only the social comparison treatment is out of boundaries by 4 gallons), and results for caliper-matching for both treatments satisfy *Criterion 1*.

[32] We only show results for single difference estimator (with OLS and IPW) and fixed-effect panel data (difference-in-differences estimator) because they were the estimators with lower bias in Section 7.

Using the trimmed sample, the performances of OLS and FEPD do not improve consistently. Furthermore, the performances differ substantially depending upon the model (Logit or Probit) used to estimate the propensity scores that determine the optimal trimming threshold. For OLS, trimming with Probit propensity scores performs better for both treatments: just changing the choice model from Logit to Probit, the percentage increases by approximately 15%. We cannot infer the same pattern for FEPD: its performance does not improve homogeneously using Probit models. Only the technical information treatment improves (41% of the times) while the social comparison treatment just improves by 2% of the time. The performance of IPW weights and trimming is also poor, providing evidence that IPW performance depends upon small changes in the sample, which is undesirable. The difference in performance in our study based on whether trimming is done with Logit or Probit models might be exacerbated because of the concentration of propensity scores on the extremes (closer to zero and one). Both choice models differ more at the tails than in the middle range of the distribution (e.g., Logistic models have slightly flatter tails, i.e. the Probit model approaches to the axes more quickly than the Logit model) (Xie and Manski, 1988).

Using the matching sample, however, OLS and FEPD improve their performances. Nevertheless, OLS has mixed results depending upon the type of matching. OLS with matching (alone) recovers 60% times the experimental social comparison treatment (specifically, within its 95% CI), but with matching-caliper that percentage decreases to 48%. Instead, for the technical information treatment, matching (alone) recovers only 14%, but caliper-matching increases up to 70%. OLS results are thus highly sensitive to the covariates included and the sample selection.

FEPD, on the contrary, improves its performance substantially when using matching: it increases from nearly zero under full sample to 16% and 50% for the social comparison and technical information treatments, respectively. Caliper-matching expands further the achievement of FEPD: 73% and 79% of the time the social comparison and technical information treatments estimates, respectively, are within the experimental benchmark's 95% CI. Caliper-matching sample combined with FEPD is the only approach that satisfies *Criterion 2*'s consistency prerequisite under bootstrapping (i.e. both treatments achieve similar conclusions with similar percentage rates). We believe this achievement arises because pre-processing the data with matching methods makes panel data identifying assumptions more plausible and, therefore, the treatment effect estimates from the non-experimental design more accurate.[33]

These results are illustrated graphically in Figure 10, Figure 11 and Figure 12. Figure 10 shows results for OLS with bootstrapping for the full sample and the matched sample (with and without caliper). The dependent variable is summer 2007 and the experimental benchmarks are -1.70 and -0.13 thousands of gallons reduction for the social comparison and technical advice treatment, respectively. These two experimental benchmark estimates are represented as vertical dashed lines. Figure 10 suggests that non-experimental OLS results improve substantially with the matched sample compared to the full sample. Further, OLS with the matched sample has the distribution with the lowest variance. In the case of the technical advice treatment, on the right side, the caliper-matching sample outperforms the sample using matching without calipers. In the case of the social comparison treatment, on the left side of Figure 10, the performances of the two samples are similar.

Figure 11 shows OLS results (with and without IPW weights) for the trimmed sample. OLS results without weights have large variance and do not perform well for both treatments. Using IPW weights, either ATE or ATT from Logit and Probit models, results improve, and distributions look more alike, but not in a consistent way for both treatments.

Likewise, Figure 12 shows results for FEPD with bootstrapping. In this case, the dependent variable is monthly water consumption and the experimental benchmarks are -0.353 and -0.006 thousands of gallons reduction for the social comparison and technical advice treatment, respectively. In this case it is clear that the caliper-matching sample outperforms, for both treatments, the other samples. Thus, the non-experimental panel data, fixed-effect estimator combined with matching to pre-process the data does a good job of replicating the experimental benchmark.

[33] Appendix 15 shows results not only based on the point estimate. It also includes the significance level (i.e. significant for the social comparison treatment and insignificant for the technical information treatment). Percentages are similar to those showed in Table 27.

11. Conclusions

Design-replication studies assess the ability of non-experimental designs to replicate impact estimates from experimental designs. If the non-experimental estimate is close to the experimental estimate, the non-experimental design is labeled as "successful." Our design-replication study builds on a randomized field experiment that was used to test the effectiveness of non-pecuniary messages to induce voluntary reductions in water consumption. We focus on two experimental treatments: (i) a technical information message, which gave households technical information on how to reduce water consumption, and (ii) a social comparison message treatment, in which households received the technical information message augmented by social norm-based encouragement and a social comparison in which own consumption is compared to median consumption in the county. In the randomized experimental trial, the technical information did not have statistically significant, or economically relevant, impact on residential household water consumption. In contrast, the social comparison treatment had a policy-relevant and statistically significant ($p<0.01$) estimated treatment effect.

Using the treatment groups and a non-experimental comparison group from a neighboring county, we estimate treatment effects of the two treatments using popular econometric approaches. We find that when using standard panel data approaches with fixed effects estimators, the estimates and inferences from the non-experimental data are only similar to those from the experimental data if the data are pre-processed via matching methods. With caliper-matching followed by a fixed-effect regression model, the non-experimental estimator does a good job of replicating the experimental benchmark. Importantly, the inferences drawn from a test of the null hypothesis of zero impact are the same using experimental and non-experimental estimators.

Furthermore, the results using panel data show that: (1) panel data are not a panacea for addressing bias (even time-invariant, unobservable bias) and (2) careful consideration of the validity of the identifying assumptions in any effort to identify causal impacts is critical (Morgan and Winship, 2007; Angrist and Pischke 2009). Pre-processing the data with matching methods makes the identifying assumptions more plausible and the treatment effect estimates from the non-experimental design more accurate. In a sense, pre-processing via matching followed by estimation in a fixed effect model is similar in spirit to an OLS model that includes both lagged dependent variables and unobserved household fixed effects. If what makes Cobb County households different from Fulton County households are time-invariant household unobservables and time-varying unobservables that are captured by household water use h periods ago, the estimation procedure we follow can account for both sources of bias and permit consistent estimation of the treatment effect. In contrast, consistent estimation of the parameters in an OLS model that includes both lagged dependent variables and unobserved household fixed effects requires the use of Y_{it-2} as an instrument for Y_{it-1}, which requires the strong (and untenable in our context) assumption that $Y_{it-2} \perp \varepsilon_{it}$.

Finally, we also find that without pre-processing the data to ensure the selection of an appropriate counterfactual, the estimates and inferences from a fixed-effects panel data estimator can be quite different from the experimental benchmark: the non-experimental estimates of the social comparison treatment effect are almost three times larger than the experimental estimates, while the non-experimental experimental estimates of the technical information treatment effect are over ten times larger. Discarding observations with high and low propensity score (trimming) improve partially the non-experimental estimates. Adding Inverse Probability Weights to the trimmed sample also improves the non-experimental estimates, but not substantially; none of these approaches to reweighting the data before performing the parametric analysis performs as well as covariate matching.

The implication of our study is that under simpler conditions, i.e., no selection bias due to the treatment, the non-experimental methods used still fail to recover experimental estimates. Thus, social scientists should applied non-experimental methods cautiously, recognizing their advantages, disadvantages, assumptions, and –more importantly– acknowledging clearly which method suits better for the specific data characteristics.

12. References

Abadie, Alberto and Javier Gerdeazabal (2003), The Economic Costs of Conflict: A Case Study of the Basque Country, *The American Economic Review*, Vol. 93, No. 1, pp. 113–132.

Abadie, Alberto and Guido Imbens (2006), Large Sample Properties of Matching Estimators for Average Treatment Effects, *Econometrica*, Vol. 74**,** pp. 235–267.

Agodini, Roberto and Mark Dynarski (2004), Are Experiments the Only Option? A Look at Dropout Prevention Programs, *The Review of Economics and Statistics*, Vol. 86, No. 1, pp. 180–194.

Angrist, Joshua and Alan Krueger (1999), Empirical Strategies in Labor Economics, In: [Ed.] Orley Ashenfelter and David Card, Handbook of Labor Economics, Vol. 3, Chapter 23, pp. 1277–1366.

Angrist, Joshua and Jörn-Steffen Pischke (2009), Mostly Harmless Econometrics: An Empiricist's Companion, Princeton University Press.

Arceneaux, Kevin; Alan Gerber and Donald Green (2006). Comparing Experimental and Matching Methods Using a Large-Scale Voter Mobilization Experiment. *Political Analysis,* Vol. 14, pp. 37–62.

Besley, Timothy, and Robin Burgess (2004), Can Labor Regulation Hinder Economic Performance? Evidence from India, *Quarterly Journal of Economics*, Vol. 113, pp. 91–134.

Blundell, Richard and Monica Costa-Dias (2009), Alternative Approaches to Evaluation in Empirical Microeconomics, *Journal of Human Resources*, Vol. 44, No. 3, pp. 565–640.

Busso, Matias; John DiNardo; Justin McCrary (2009), Finite Sample Properties of Semiparametric Estimators of Average Treatment Effects, Working Paper.

Busso, Matias; John DiNardo; Justin McCrary (2011), New Evidence on the Finite Sample Properties of Propensity Score Reweighting and Matching Estimators, Working Paper.

Card, David and Alan Krueger (1994), Minimum Wages and Employment: a Case Study of the Fast-Food Industry in New Jersey and Pennsylvania. *The American Economic Review*, Vol. 84, pp. 772–793.

Crump, Richard; Joseph Hotz; Guido Imbens; Oscar Mitnik (2009), Dealing with Limited Overlap in Estimation of Average Treatment Effects, *Biometrika*, Vol. 96, No. 1, pp. 187–199.

Davis, Lucas (2004), The Effect of Health Risk on Housing Values: Evidence from a Cancer Cluster, *American Economic Review*, Vol. 94, No. 5, pp. 1693–1704.

Dehejia, Rajeev and Sadek Wahba (1999), Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs, *Journal of the American Statistical Association*, Vol. 94, No. 448, pp. 1053–1062.

Dehejia, Rajeev and Sadek Wahba (2002), Propensity Score-Matching Methods for Nonexperimental Causal Studies, *The Review of Economics and Statistics*, Vol. 84, No. 1, pp. 151–161.

Diaz, Juan Jose and Sudhanshu Handa (2006), An Assessment of Propensity Score Matching as a Nonexperimental Impact Estimator: Evidence from Mexico's PROGRESA Program, *The Journal of Human Resources*, Vol. 41, No. 2, pp. 319–345.

Drake, Christiana (1993), Effects of Misspecification of the Propensity Score on Estimators of Treatment Effect, *Biometrics*, Vol. 49, No. 4, pp. 1231–1236.

Emsley, Richard; Graham Dunn; Mark Lunt; Andrew Pickles (2008), Implementing double-robust estimators of causal effects, *The Stata Journal*, Vol. 8, No. 3, pp. 334–353.

Ferraro, Paul and Michael Price (forthcoming), Using non-pecuniary strategies to influence behavior: Evidence from a large-scale field experiment, *The Review of Economics and Statistics*.

Ferraro, Paul and Juan Jose Miranda (2011), Heterogeneous Treatment Effects and Mechanisms in Information-Based Environmental Policies: Evidence from a Large-Scale Field Experiment, Working Paper.

Ferraro, Paul, Juan Jose Miranda and Michael Price (2011), The Persistence of Treatment Effects with Non-Pecuniary Policy Instruments: Evidence from a Randomized Environmental Policy Experiment, *The American Economic Review: Papers and Proceedings*, Vol. 101, No. 3, pp. 318–322.

Fraker, Thomas and Rebecca Maynard (1987), The Adequacy of Comparison Group Designs for Evaluations of Employment-Related Programs, *Journal of Human Resources*, Vol. 22, No. 2, pp. 194–227.

Galiani, Sebastian; Paul Gertler and Ernesto Schargrodsky (2005), "Water for Life: The Impact of the Privatization of Water Services on Child Mortality", *Journal of Political Economy*, Vol. 113, No. 1, pp. 83–120.

Gibbons, Charles; Juan Carlos Suarez; Michael Urbancic (2011), Broken or Fixed Effects?, Working Paper, University of California at Berkeley.

Glazerman, Steven; Dan Levy and David Myers (2003). Nonexperimental versus Experimental Estimeates of Earnings Impacts. *The Annals of the American Academy of Political and Social Science*, Vol. 589, No. 1, pp. 63–93.

Greenberg, David; Charles Michalopoulos; Philip Robins (2006), Do Experimental and Nonexperimental Evaluations Give Different Answers about the Effectiveness of Government-Funded Training Programs?, *Journal of Policy Analysis and Management*, Vol. 25, No. 3, pp. 523–552.

Greenstone, Michael and Ted Gayer (2009), Quasi-Experimental and Experimental Approaches to Environmental Economics, *Journal of Environmental Economics and Management*, Vol. 57, pp. 21–44.

Hahn, Jinyong (1998), On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects, *Econometrica*, Vol. 66, pp. 315–331.

Handa, Sudhanshu and John Maluccio (2006), Matching the Gold Standard: Comparing Experimental and Nonexperimental Evaluation Techniques for a Geographically Targeted Program, *Economic Development and Cultural Change*, Vol. 58, April 2010, pp. 415–447.

Heckman, James; Hidehiko Ichimura; Petra Todd (1997). Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Program. *Review of Economic Studies*, Vol. 64, No. 4, pp. 605–654.

Heckman, James; Hidehiko Ichimura; Jeffrey Smith; Petra Todd (1998a). Characterizing Selection Bias Using Experimental Data. *Econometrica*, Vol. 55, No. 5, pp. 1017–1098.

Heckman, James; Hidehiko Ichimura; Petra Todd (1998b). Matching as an Econometric Evaluation Estimator. *Review of Economic Studies*, Vol. 65, No. 2, pp. 261–294.

Hill, Jennifer; Jerome Reiter; Elaine Zanutto (2004), A Comparison of Experimental and Observational Data Analyses, In: "Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives", Andrew Gelman and Xiao-Li Meng [Ed.]; New York: Wiley.

Hirano, Keisuke; Guido Imbens; Geert Ridder (2003), Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score, *Econometrica*, Vol. 71, No. 4, pp. 1161–1189.

Ho, Daniel; Kosuke Imai; Gary King; Elizabeth Stuart (2007), Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference, *Political Analysis*, Vol. 15, pp. 199–236.

Hotz, Joseph; Guido Imbens and Julie Mortimer (2005), Predicting the efficacy of future training programs using past experiences at other locations, *Journal of Econometrics*, Vol. 125, pp. 241–270.

Huber, Martin; Michael Lechner; Conny Wunsch (2010), How to Control for Many Covariates? Reliable Estimators based on the Propensity Score, University of St. Gallen, Department of Economics, Discussion Paper No. 2010-30.

Imai, Kosuke and In Song Kim (2011), Understanding and Improving Linear Fixed Effects: Regression Models for Causal Inference, Working Paper, Princeton University.

Imbens, Guido and Jeffrey Wooldridge (2009), Recent Developments in the Econometrics of Program Evaluation, *Journal of Economic Literature*, Vol. 47, No. 11, pp. 5–86.

Lalonde, Robert (1986), Evaluating the econometric evaluations of training with experimental data, *The American Economic Review*, Vol. 76, pp. 604–620.

Lalonde, Robert and Rebecca Maynard (1987), How Precise are Evaluations of Employment and Training Programs – Evidence from a Field Experiment, *Evaluation Review*, Vol. 11, No. 4, pp. 428–451.

Lee, Wang-Sheng (2011), Propensity Score Matching and Variations on the Balancing Test, *Empirical Economics*, 26 May 2011, pp. 1–34.

Lee, David and Thomas Lemieux (2010), Regression Discontinuity Designs in Economics, *Journal of Economic Literature*, Vol. 48, June, pp. 281–355.

Millimet, Daniel and Rusty Tchernis (2009), On the Specification of Propensity Scores, With Applications to the Analysis of Trade Policies, Journal of Business & Economic Statistics, Vol. 29, No. 3, pp. 397–414.

Morgan, Stephen and Christopher Winship (2007), Counterfactuals and Causal Inference: Methods and Principles for Social Research, Cambridge University Press.

McKenzie, David; John Gibson; Steven Stillman (2010), How Important Is Selection? Experimental vs. Non-Experimental Measures of the Income Gains from Migration, *Journal of the European Economic Association*, Vol. 8, No. 4, pp. 913–945.

Ravallion, Martin (2008), Evaluating Anti-Poverty Programs, In: [Ed.] T. Paul Schultz and John Strauss, Handbook of Development Economics, Vol. 4, Chapter 59, pp. 3787–3846.

Rosenbaum, Paul and Donald Rubin (1983). The Central Role of the Propensity Score in Observational Studies for Casual Effects. *Biometrika*, Vol. 70, No. 1, pp. 41–55.

Rosenbaum, Paul and Donald Rubin (1985), Constructing a Control Group Using Multivariate Matched Sampling Methods that Incorporate the Propensity Score, *The American Statistician*, Vol. 39, No. 1, pp. 33–38.

Sekhon, Jasjeet (2010), Package 'Matching'. R documentation, University of California at Berkeley, available at: http://sekhon.berkeley.edu/matching/Match.html

Sekhon, Jasjeet (2011), Multivariate and Propensity Score Matching Software with Automated Balance Optimization: The Matching Package for R, *Journal of Statistical Software*, Vol. 42, No. 7, pp. 1–52.

Shadish, William, M.H. Clark, Peter Steiner (2008), Can Nonrandomized Experiments Yield Accurate Answers? A Randomized Experiment Comparing Random and Nonrandom Assignments, *Journal of the American Statistical Association*, Vol. 103, No. 484, pp. 1334–1356.

Smith, Jeffrey and Arthur Sweetman (2009), Putting the Evidence in Evidence Based Policy, Chapter 4, In: Strengthening Evidence-based Policy in the Australian Federation – Roundtable Proceedings, available at: http://www.pc.gov.au/research/confproc/strengthening-evidence

Smith, Jeffrey and Petra Todd (2005), Does matching overcome Lalonde's critique of nonexperimental estimators?, *Journal of Econometrics,* Vol. 125, pp. 305–353.

Stuart, Elizabeth (2009), Matching methods for causal inference: A review and a look forward, Department of Biostatistics, John Hopkins University, Working Paper.

Swaffield, Joanna (2001), Does measurement error bias fixed-effects estimates of the union wage effect? *Oxford Bulletin of Economics and Statistics*, Vol. 63, No. 4, pp. 437–457.

Train, Kenneth (2009), Discrete Choice Methods with Simulations, Cambridge University Press, 2nd Edition.

Wilde, Elizabeth and Robinson Hollister (2007), How Close Is Close Enough? Evaluating Propensity Score Matching Using Data from a Class Size Reduction Experiment, *Journal of Policy Analysis and Management*, Vol. 26, No. 3, pp. 455–477.

Wooldridge, Jeffrey (2005), Fixed-Effects and Related Estimators for Correlated Random-Coefficient and Treatment Effect Panel Data Models, *The Review of Economics and Statistics*, Vol. 87, No. 2, pp. 385–390.

Xie, Yu and Charles Manski (1988), The Logit Model, the Probit Model, and Response-Based Samples, Working Paper, University of Wisconsin – Madison.

Table 9
Summary of Non-Experimental Methods and Sample Selection

	Sample			
	Full	Trimmed	Matching	Caliper Matching
- Simple Difference	✓	✓	✓*	✓*
- Simple Difference with Covariate Matching and Bias Adjustment			✓	✓
- Simple Difference with OLS	✓	✓	✓	✓
- Simple Difference with IPW		✓	✓	✓
- Difference-in-Differences	✓	✓	✓**	✓**
- Difference-in-Differences with Covariate Matching and Bias Adjustment			✓	✓
- Difference-in-Differences with OLS	✓	✓	✓	✓
- Difference-in-Differences with IPW		✓	✓	✓
- Fixed-effects Panel Data Estimator	✓	✓	✓	✓

* Simple Difference with Matching corresponds to Covariate Matching method without bias adjustment.

** Difference-in-Differences with Matching corresponds to Covariate Matching method without bias adjustment.

Table 10
Descriptive Statistics: Cobb and Fulton County

	Cobb County	Fulton County
Population	607,751	816,006
Urban Population	99.5%	97.9%
Race: White alone	72.4%	48.1%
Race: African American alone	18.8%	44.6%
Households	227,487	321,242
Race: White alone	75.3%	53.5%
Race: African American alone	18.3%	41.1%
Average Household Size	2.64	2.44
Housing Units	237,522	348,632
Urban Units	99.5%	98.1%
Vacant Units	4.2%	7.9%
Population 25 years and over	395,349	527,738
Education ≥ College	68.0%	64.6%
Income Distribution		
Less than $24,999	24,427	70,291
$20,000 to $39,999	45,250	69,098
$40,000 to $59,999	47,340	52,047
$60,000 to $99,999	61,986	60,644
$100,000 or more	48,587	69,186
Per capita income 1999	27,863	30,003
Median household income 1999	58,289	47,321
Population below Poverty Level[1/]	6.5%	15.7%

[1/] Compared with 1999 income.

Source: 2000 US Census.

Table 11
Water Consumption by Seasons*

Variable	Indicator	(1)	(2)	(3)	(4)
		Cobb County			Fulton County
		Technical Information Treatment	Social Comparison Treatment	Experimental Control	Non-Experimental Control
Summer 2006 1/	Mean	39.04	39.03	38.95	48.94
	St. Dev.	28.48	29.21	29.90	44.45
Summer 2007 1/	Mean	36.03	34.53	36.04	41.55
	St. Dev.	30.30	26.13	29.01	52.64
Spring 2007 2/	Mean	27.14	26.67	27.20	25.19
	St. Dev.	19.78	18.91	44.94	71.16
Winter 2006/2007 3/	Mean	25.25	25.19	25.19	27.78
	St. Dev.	14.59	14.62	16.05	94.01

* In thousands of gallons.

1/ Summer season includes June, July, August, September.

2/ Spring season includes March, April, May.

3/ Winter season include from November, December, January, February.

Source: Cobb County Water System and Fulton County Water Service Division.

Table 12
Descriptive Statistics: Cobb and Fulton County
Household and Neighborhood Variables

Variable	Indicator	(1)	(2)	(3)	(4)
		Cobb County			Fulton County
		Technical Information Treatment	Social Comparison Treatment	Experimental Control	Non-Experimental Control
Tax Assessor (Household) Variables					
Fair Market Value ($)	Mean	249,469.30	252,355.80	250,604.30	350,238.30
	St. Dev.	157,745.30	174,657.30	161,127.80	219,111.70
Age of Home (Years)	Mean	22.04	22.04	22.04	16.99
	St. Dev.	12.74	12.87	12.97	8.65
Size of Property (Acres)	Mean	0.18	0.20	0.20	0.62
	St. Dev.	1.23	1.07	1.17	0.90
Census (Neighborhood) Variables 1/					
% of People with Higher Degree	Mean	0.73	0.73	0.73	0.85
	St. Dev.	0.15	0.15	0.15	0.07
% of People Below Poverty Level	Mean	0.04	0.04	0.04	0.03
	St. Dev.	0.04	0.04	0.04	0.03
Per-capita Income	Mean	30,813.86	30,817.17	30,804.28	42,383.66
	St. Dev.	9,220.59	9,218.93	9,195.02	10,537.61
% Renter-Occupied Housing Units	Mean	0.12	0.12	0.12	0.16
	St. Dev.	0.15	0.15	0.15	0.16
% White	Mean	0.84	0.84	0.84	0.87
	St. Dev.	0.13	0.13	0.13	0.06

1/ At census Block group level.

Source: 2000 US Census, Cobb County and Fulton County Tax Assessor.

Table 13
OLS Regression: Experimental Results
Dependent Variable: Water Use on Summer 2007*
(with meter route fixed effects)

	(1)	(2)	(3)
Technical Information Treatment	-0.0132	-0.149	-0.124
	(0.303)	(0.205)	(0.204)
Social Comparison Treatment	-1.534***	-1.678***	-1.732***
	(0.267)	(0.178)	(0.178)
Water Use from June - November 2006		0.357***	0.341***
		(0.0166)	(0.0162)
Water Use in April and May 2007		0.824***	0.812***
		(0.0576)	(0.0580)
Fair Market Value			1.82e-05***
			(3.90e-06)
Age of Home			0.0268*
			(0.0142)
Size of Property (Acres)			0.0761
			(0.126)
% of People with Higher Degree			1.274
			(1.854)
% of People Below Poverty Level			-3.774
			(4.266)
Per-capita Income			4.11e-05
			(4.06e-05)
% of Renter-Occupied Housing Units			0.393
			(1.068)
% White			-0.284
			(1.698)
Constant	34.08***	1.284	-4.397
	(2.301)	(1.883)	(2.741)
Observations	82,027	82,027	81,585
R-squared	0.129	0.632	0.637

*In thousands of gallons.
Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1
Source: Cobb County Water System, Tax Assessor and 2000 US Census.

Table 14
Panel Data Regression with Fixed Effects: Experimental Results
Dependent Variable: Monthly Water Use*

	(1)
Technical Information Treatment * Post Experiment [1/]	-0.00573
	(0.0539)
Social Comparison Treatment * Post Experiment [1/]	-0.353***
	(0.0461)
Post Experiment	0.561***
	(0.0219)
Constant	8.446***
	(0.00430)
Observations	1,394,455
Number of code	82,027
R-squared	0.000

*In thousands of gallons. Regression analysis period is from May 2006 to September 2007.

1/ Post Experiment represents from June 2007 to September 2007.

Robust standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

Source: Cobb County Water System.

Table 15
Social Comparison Treatment: Balance Test on Covariates

	Full Sample	Trimmed Sample	Matching without Calipers	Matching with Calipers	% Improvement (1) to (2)	% Improvement (1) to (3)	% Improvement (1) to (4)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Water Use from June - November 2006						
Mean Difference	-8.565	-4.416	-2.480	-0.697	48%	71%	92%
Standardized Mean Diff.	-21.170	-11.498	-6.130	-1.919	46%	71%	91%
Mean Raw eQQ Diff.	8.667	4.524	2.917	1.825	48%	66%	79%
Variance Ratio (Tr./Ct.)	0.551	0.538	1.102	1.082	-3%	77%	82%
	Water Use in April and May 2007						
Mean Difference	0.875	1.088	2.082	2.419	-24%	-138%	-176%
Standardized Mean Diff.	7.629	9.856	18.151	23.128	-29%	-138%	-203%
Mean Raw eQQ Diff.	2.334	3.542	2.136	2.442	-52%	8%	-5%
Variance Ratio (Tr./Ct.)	0.033	0.015	1.722	1.770	-2%	25%	20%
	Fair Market Value						
Mean Difference	-96,280	-58,738	-45,676	-26,950	39%	53%	72%
Standardized Mean Diff.	-55.1	-38.2	-26.1	-18.9	31%	53%	66%
Mean Raw eQQ Diff.	97,173	58,755	47,582	28,213	40%	51%	71%
Variance Ratio (Tr./Ct.)	0.656	0.675	1.192	1.138	6%	44%	60%
	Age of Home						
Mean Difference	5.060	1.050	1.850	0.533	79%	63%	89%
Standardized Mean Diff.	39.319	10.061	14.377	5.345	74%	63%	86%
Mean Raw eQQ Diff.	5.294	1.845	1.952	0.885	65%	63%	83%
Variance Ratio (Tr./Ct.)	2.213	1.333	1.397	1.216	73%	67%	82%
	Size of Property (Acres)						
Mean Difference	-0.419	-0.370	-0.311	-0.305	12%	26%	27%
Standardized Mean Diff.	-40.983	-102.130	-30.372	-66.572	-149%	26%	-62%
Mean Raw eQQ Diff.	0.460	0.400	0.328	0.314	13%	29%	32%
Variance Ratio (Tr./Ct.)	1.326	1.787	1.020	1.194	-141%	94%	40%
	% of People with Higher Degree						
Mean Difference	-0.126	-0.056	-0.069	-0.027	56%	45%	79%
Standardized Mean Diff.	-86.258	-55.861	-47.145	-29.831	35%	45%	65%
Mean Raw eQQ Diff.	0.126	0.057	0.069	0.028	55%	45%	78%
Variance Ratio (Tr./Ct.)	4.275	1.456	1.953	1.397	86%	71%	88%
	% of People Below Poverty Level						
Mean Difference	0.005	0.001	0.007	0.002	84%	-32%	70%
Standardized Mean Diff.	15.567	2.950	20.571	7.423	81%	-32%	52%
Mean Raw eQQ Diff.	0.008	0.004	0.009	0.004	49%	-14%	56%
Variance Ratio (Tr./Ct.)	1.892	1.101	1.145	1.082	89%	84%	91%
	Per-capita Income						
Mean Difference	-11,538	-5,259	-5,255	-3,105	54%	54%	73%
Standardized Mean Diff.	-125.3	-61.6	-57.0	-37.8	51%	54%	70%
Mean Raw eQQ Diff.	11,537	5,686	5,395	3,516	51%	53%	70%
Variance Ratio (Tr./Ct.)	0.764	1.468	1.378	1.330	-99%	-61%	-40%
	% of Renter-Occupied Housing Units						
Mean Difference	-0.036	-0.018	0.007	-0.001	49%	80%	97%
Standardized Mean Diff.	-24.813	-16.692	5.011	-1.061	33%	80%	96%
Mean Raw eQQ Diff.	0.039	0.022	0.016	0.010	42%	57%	73%
Variance Ratio (Tr./Ct.)	0.823	0.616	0.899	1.033	-117%	43%	81%
	% White						
Mean Difference	-0.027	-0.004	-0.041	-0.005	85%	-53%	80%
Standardized Mean Diff.	-20.323	-3.649	-31.030	-11.160	82%	-53%	45%
Mean Raw eQQ Diff.	0.040	0.031	0.041	0.009	22%	-1%	77%
Variance Ratio (Tr./Ct.)	4.061	2.496	3.172	0.811	51%	29%	94%

Source: Cobb County Water System, Fulton County Water Service Division, Tax Assessor and 2000 US Census.

Table 16
Technical Information Treatment: Balance Test on Covariates

	Full Sample	Trimmed Sample	Matching without Calipers	Matching with Calipers	% Improvement (1) to (2)	% Improvement (1) to (3)	% Improvement (1) to (4)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Water Use from June - November 2006						
Mean Difference	-8.786	-4.021	-2.495	-0.861	54%	72%	90%
Standardized Mean Diff.	-22.798	-10.383	-6.474	-2.360	54%	72%	90%
Mean Raw eQQ Diff.	8.887	4.135	2.766	1.736	53%	69%	80%
Variance Ratio (Tr./Ct.)	0.500	0.559	1.057	1.061	12%	89%	88%
	Water Use in April and May 2007						
Mean Difference	0.902	1.351	2.154	2.585	-50%	-139%	-187%
Standardized Mean Diff.	7.788	11.321	18.594	23.298	-45%	-139%	-199%
Mean Raw eQQ Diff.	2.363	3.742	2.219	2.610	-58%	6%	-10%
Variance Ratio (Tr./Ct.)	0.033	0.018	1.847	1.869	-2%	12%	10%
	Fair Market Value						
Mean Difference	-99,257	-57,902	-46,502	-27,024	42%	53%	73%
Standardized Mean Diff.	-63.0	-38.6	-29.5	-19.8	39%	53%	69%
Mean Raw eQQ Diff.	100,137	57,980	48,010	28,746	42%	52%	71%
Variance Ratio (Tr./Ct.)	0.532	0.658	1.101	1.163	27%	78%	65%
	Age of Home						
Mean Difference	5.064	1.223	1.955	0.629	76%	61%	88%
Standardized Mean Diff.	39.743	11.632	15.344	6.265	71%	61%	84%
Mean Raw eQQ Diff.	5.282	1.965	2.040	0.919	63%	61%	83%
Variance Ratio (Tr./Ct.)	2.169	1.370	1.410	1.226	68%	65%	81%
	Size of Property (Acres)						
Mean Difference	-0.433	-0.359	-0.314	-0.309	17%	27%	28%
Standardized Mean Diff.	-36.992	-110.530	-26.876	-76.632	-199%	27%	-107%
Mean Raw eQQ Diff.	0.464	0.385	0.337	0.319	17%	28%	31%
Variance Ratio (Tr./Ct.)	1.735	1.676	1.224	1.293	8%	70%	60%
	% of People with Higher Degree						
Mean Difference	-0.126	-0.058	-0.068	-0.026	54%	46%	79%
Standardized Mean Diff.	-85.732	-57.345	-46.536	-29.532	33%	46%	66%
Mean Raw eQQ Diff.	0.126	0.059	0.069	0.028	53%	46%	78%
Variance Ratio (Tr./Ct.)	4.322	1.494	1.930	1.353	85%	72%	89%
	% of People Below Poverty Level						
Mean Difference	0.006	0.001	0.007	0.002	83%	-24%	68%
Standardized Mean Diff.	16.814	3.433	20.863	8.491	80%	-24%	50%
Mean Raw eQQ Diff.	0.008	0.004	0.009	0.004	49%	-11%	54%
Variance Ratio (Tr./Ct.)	1.935	1.116	1.133	1.093	88%	86%	90%
	Per-capita Income						
Mean Difference	-11,533	-5,223	-5,184	-3,028	55%	55%	74%
Standardized Mean Diff.	-125.2	-60.9	-56.3	-37.0	51%	55%	70%
Mean Raw eQQ Diff.	11,532	5,650	5,328	3,462	51%	54%	70%
Variance Ratio (Tr./Ct.)	0.765	1.495	1.397	1.341	-111%	-69%	-45%
	% of Renter-Occupied Housing Units						
Mean Difference	-0.034	-0.021	0.009	0.000	38%	74%	99%
Standardized Mean Diff.	-22.388	-18.744	5.844	-0.424	16%	74%	98%
Mean Raw eQQ Diff.	0.036	0.024	0.017	0.011	33%	53%	71%
Variance Ratio (Tr./Ct.)	0.874	0.626	0.926	1.046	-197%	41%	64%
	% White						
Mean Difference	-0.027	-0.004	-0.040	-0.006	86%	-49%	80%
Standardized Mean Diff.	-20.497	-3.272	-30.510	-11.812	84%	-49%	42%
Mean Raw eQQ Diff.	0.040	0.033	0.041	0.010	19%	0%	76%
Variance Ratio (Tr./Ct.)	4.189	2.691	3.169	0.797	47%	32%	94%

Source: Cobb County Water System, Fulton County Water Service Division, Tax Assessor and 2000 US Census.

Table 17
Non-Experimental Results: Single Difference and Difference-in-Differences for Social Comparison Treatment*

	(1)	(2)	(3)	(4)	(5)	(6)
	Social Comparison Treatment					
	Cobb Treated Group		Fulton Control Group		Diff. in Diff.	Single Diff.
	Summer 2007	Summer 2006	Summer 2007	Summer 2006		
- Full Sample	34.53	39.03	41.55	48.94	2.89	-7.03
	(0.261)	(0.292)	(0.299)	(0.253)	(0.343)	(0.397)
- Trimmed Sample	35.43	40.71	39.70	46.10	1.13	-4.26
	(0.294)	(0.332)	(0.448)	(0.299)	(0.555)	(0.603)
- Matched Sample (no calipers)	34.54	39.06	36.42	43.46	2.52	-1.88
	(0.261)	(0.293)	(0.287)	(0.307)	(0.295)	(0.388)
- Matched Sample (calipers)	35.56	40.25	36.21	43.00	2.09	-0.65
	(0.289)	(0.320)	(0.301)	(0.325)	(0.318)	(0.417)
- Matched Sample (no calipers) & Bias Correction					0.65	-1.10
					(1.021)	(0.945)
- Matched Sample (calipers) & Bias Correction					0.66	-1.12
					(0.723)	(0.684)
- Trimmed Sample & IPW (ATT)	35.43	40.71	38.38	44.75	1.10	-2.94
	(0.299)	(0.344)	(1.032)	(1.028)	(0.590)	(1.074)
- Matched Sample (no calipers) & IPW (ATT)	35.43	40.71	31.09	37.61	1.24	4.34
	(0.299)	(0.344)	(0.283)	(0.329)	(0.457)	(0.610)
- Matched Sample (calipers) & IPW (ATT)	35.91	41.24	30.96	37.37	1.07	4.94
	(0.317)	(0.359)	(0.258)	(0.315)	(0.461)	(0.573)
- Trimmed Sample & IPW (ATE)	37.11	43.11	39.38	45.78	0.40	-2.27
	(0.433)	(0.499)	(0.531)	(0.410)	(0.568)	(0.685)
- Matched Sample (no calipers) & IPW (ATE)	37.11	43.11	33.24	39.94	0.70	3.88
	(0.433)	(0.499)	(0.259)	(0.293)	(0.476)	(0.639)
- Matched Sample (calipers) & IPW (ATE)	37.79	43.80	33.02	39.58	0.55	4.77
	(0.466)	(0.529)	(0.260)	(0.300)	(0.472)	(0.616)

*In thousands of gallons.
Standard errors in parentheses.
Source: Cobb County Water System and Fulton County Water Service Division.

Table 18
Non-Experimental Results: Single Difference and Difference-in-Differences for Technical Information Treatment*

	(1)	(2)	(3)	(4)	(5)	(6)
	Technical Information Treatment					
	Cobb Treated Group		Fulton Control Group		Diff. in Diff.	Single Diff.
	Summer 2007	Summer 2006	Summer 2007	Summer 2006		
- Full Sample	36.03	39.04	41.55	48.94	4.38	-5.52
	(0.302)	(0.283)	(0.299)	(0.253)	(0.357)	(0.425)
- Trimmed Sample	37.48	40.65	38.86	45.61	3.58	-1.38
	(0.354)	(0.325)	(0.437)	(0.293)	(0.391)	(0.486)
- Matched Sample (no calipers)	36.03	39.02	36.10	43.16	4.08	-0.06
	(0.302)	(0.283)	(0.266)	(0.296)	(0.305)	(0.402)
- Matched Sample (calipers)	37.49	40.48	36.42	43.10	3.69	1.07
	(0.334)	(0.324)	(0.292)	(0.326)	(0.327)	(0.443)
- Matched Sample (no calipers) & Bias Correction					3.07	1.61
					(0.700)	(0.672)
- Matched Sample (calipers) & Bias Correction					2.02	0.56
					(0.512)	(0.508)
- Trimmed Sample & IPW (ATT)	37.48	40.65	37.36	44.21	3.68	0.12
	(0.379)	(0.343)	(0.757)	(0.819)	(0.623)	(0.847)
- Matched Sample (no calipers) & IPW (ATT)	37.48	40.64	31.00	37.44	3.27	6.47
	(0.379)	(0.343)	(0.270)	(0.322)	(0.475)	(0.609)
- Matched Sample (calipers) & IPW (ATT)	37.85	41.21	31.60	37.96	3.01	6.25
	(0.368)	(0.358)	(0.267)	(0.331)	(0.482)	(0.593)
- Trimmed Sample & IPW (ATE)	39.38	42.99	38.51	45.28	3.16	0.86
	(0.556)	(0.515)	(0.358)	(0.378)	(0.523)	(0.661)
- Matched Sample (no calipers) & IPW (ATE)	39.37	42.99	33.09	39.84	3.14	6.29
	(0.556)	(0.515)	(0.241)	(0.282)	(0.502)	(0.653)
- Matched Sample (calipers) & IPW (ATE)	39.75	43.62	33.49	40.09	2.73	6.26
	(0.564)	(0.532)	(0.257)	(0.305)	(0.510)	(0.648)

*In thousands of gallons.
Standard errors in parentheses.
Source: Cobb County Water System and Fulton County Water Service Division.

Table 19
Non-Experimental Results: OLS Regression with Single Difference for Social Comparison Treatment
Dependent Variable: Summer 2007*

	(1)	(2)	(3)	(4)
	Full Sample		Trimmed Sample	
Social Comparison Treatment	-2.835***	-1.549**	-2.446***	1.586
	(0.445)	(0.642)	(0.597)	(1.118)
Water Use from June - November 2006	0.466***	0.430***	0.435***	0.386***
	(0.0265)	(0.0293)	(0.0454)	(0.0506)
Water Use in April and May 2007	0.110	0.107	0.0962	0.0960
	(0.109)	(0.108)	(0.104)	(0.104)
Fair Market Value		1.53e-05***		1.81e-05***
		(4.20e-06)		(6.24e-06)
Age of Home		-0.0265		-0.0421
		(0.0199)		(0.0408)
Size of Property (Acres)		1.178***		8.042***
		(0.385)		(3.091)
% of People with Higher Degree		-15.89***		-11.36***
		(3.186)		(3.945)
% of People Below Poverty Level		3.944		-5.001
		(6.721)		(8.330)
Per-capita Income		8.35e-05***		0.000126*
		(2.89e-05)		(6.57e-05)
% of Renter-Occupied Housing Units		-0.196		3.802
		(1.028)		(3.399)
% White		18.41***		15.63***
		(3.062)		(4.056)
Constant	8.773***	-0.545	10.41***	-4.683
	(1.515)	(1.804)	(2.752)	(4.211)
Observations	41,011	40,742	21,513	21,513
R-squared	0.298	0.299	0.176	0.183

*In thousands of gallons.

1/ Trimmed sample: Propensity Score estimated using logit model. Observations outside interval [0.06, 0.94] are discarded.

Robust standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

Source: Cobb County Water System, Fulton County Water Service Division, Tax Assessor and 2000 US Census.

Cont.

	(5)	(6)	(7)	(8)
	Matched Sample			
	No Calipers		Calipers	
Social Comparison Treatment	-2.746***	-1.662***	-2.135***	-1.435***
	(0.314)	(0.307)	(0.271)	(0.316)
Water Use from June - November 2006	0.365***	0.334***	0.393***	0.365***
	(0.0195)	(0.0173)	(0.0122)	(0.0124)
Water Use in April and May 2007	0.855***	0.834***	0.720***	0.695***
	(0.0713)	(0.0739)	(0.0423)	(0.0421)
Fair Market Value		1.33e-05**		1.90e-05***
		(6.35e-06)		(2.51e-06)
Age of Home		-0.0112		0.0305*
		(0.0277)		(0.0171)
Size of Property (Acres)		0.372		0.941
		(0.321)		(0.583)
% of People with Higher Degree		-1.594		-0.954
		(2.059)		(2.849)
% of People Below Poverty Level		-18.62***		-36.18***
		(6.113)		(8.306)
Per-capita Income		3.84e-05		-6.86e-05**
		(7.12e-05)		(3.40e-05)
% of Renter-Occupied Housing Units		2.657		5.128**
		(1.653)		(2.106)
% White		4.043**		8.674**
		(1.653)		(3.495)
Constant	2.746***	-2.498	3.009***	-5.767*
	(0.857)	(1.895)	(0.438)	(3.506)
Observations	19,971	19,971	14,086	14,086
R-squared	0.604	0.611	0.596	0.605

*In thousands of gallons.

1/ Trimmed sample: Propensity Score estimated using logit model. Observations outside interval [0.06, 0.94] are discarded.

Robust standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

Source: Cobb County Water System, Fulton County Water Service Division, Tax Assessor and 2000 US Census.

Table 20
Non-Experimental Results: OLS Regression with Single Difference for Technical Information Treatment
Dependent Variable: Summer 2007*

	(1)	(2)	(3)	(4)
	Full Sample		Trimmed Sample	
Technical Information Treatment	-1.152**	0.472	0.309	3.671***
	(0.472)	(0.659)	(0.479)	(0.970)
Water Use from June - November 2006	0.478***	0.440***	0.454***	0.398***
	(0.0273)	(0.0300)	(0.0478)	(0.0524)
Water Use in April and May 2007	0.111	0.108	0.0984	0.0981
	(0.110)	(0.109)	(0.105)	(0.105)
Fair Market Value		1.96e-05***		2.80e-05***
		(4.54e-06)		(6.52e-06)
Age of Home		0.00362		0.0379
		(0.0214)		(0.0264)
Size of Property (Acres)		0.635*		5.700***
		(0.325)		(1.669)
% of People with Higher Degree		-11.96***		-6.532
		(3.219)		(4.051)
% of People Below Poverty Level		6.650		0.205
		(6.751)		(8.466)
Per-capita Income		6.32e-05**		3.96e-05
		(3.01e-05)		(5.15e-05)
% of Renter-Occupied Housing Units		-0.126		3.859
		(1.034)		(2.678)
% White		18.52***		15.25***
		(2.999)		(3.426)
Constant	7.930***	-5.579***	8.617***	-10.01***
	(1.568)	(1.980)	(2.837)	(3.735)
Observations	41,073	40,801	21,833	21,833
R-squared	0.301	0.303	0.443	0.463

*In thousands of gallons.

1/ Trimmed sample: Propensity Score estimated using logit model. Observations outside interval [0.06, 0.94] are discarded.

Robust standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

Source: Cobb County Water System, Fulton County Water Service Division, Tax Assessor and 2000 US Census.

Cont.

	(5)	(6)	(7)	(8)
	Matched Sample			
	No Calipers		Calipers	
Technical Information Treatment	-0.985***	0.161	-0.645**	-0.0616
	(0.358)	(0.324)	(0.287)	(0.321)
Water Use from June - November 2006	0.392***	0.353***	0.409***	0.379***
	(0.0263)	(0.0212)	(0.0129)	(0.0130)
Water Use in April and May 2007	0.884***	0.840***	0.798***	0.773***
	(0.0676)	(0.0692)	(0.0455)	(0.0455)
Fair Market Value		2.53e-05***		2.13e-05***
		(6.25e-06)		(2.39e-06)
Age of Home		0.0406*		0.0406**
		(0.0239)		(0.0185)
Size of Property (Acres)		0.189		0.00559
		(0.270)		(0.451)
% of People with Higher Degree		-2.486		-5.152*
		(1.980)		(2.784)
% of People Below Poverty Level		-23.53***		-48.44***
		(5.081)		(7.931)
Per-capita Income		-2.21e-05		-1.10e-05
		(6.01e-05)		(3.99e-05)
% of Renter-Occupied Housing Units		1.401		7.804***
		(1.562)		(1.781)
% White		2.816*		11.69***
		(1.487)		(3.539)
Constant	0.475	-4.169***	1.037**	-9.479***
	(0.888)	(1.605)	(0.502)	(3.322)
Observations	20,087	20,087	14,351	14,351
R-squared	0.593	0.607	0.612	0.621

*In thousands of gallons.

1/ Trimmed sample: Propensity Score estimated using logit model. Observations outside interval [0.06, 0.94] are discarded.

Robust standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

Source: Cobb County Water System, Fulton County Water Service Division, Tax Assessor and 2000 US Census.

Table 21
Non-Experimental Results: OLS Regression with Difference-in-Differences for Social Comparison Treatment
Dependent Variable: Summer 2007 – Summer 2006*

	(1)	(2)	(3)	(4)
	Full Sample		Trimmed Sample	
Social Comparison Treatment	-0.0149	0.680	-0.326	2.651**
	(0.427)	(0.646)	(0.593)	(1.049)
Water Use from June - November 2006	-0.307***	-0.335***	-0.306***	-0.340***
	(0.0197)	(0.0198)	(0.0277)	(0.0291)
Water Use in April and May 2007	0.111	0.108	0.0960	0.0960
	(0.110)	(0.109)	(0.104)	(0.103)
Fair Market Value		1.38e-05***		1.45e-05***
		(3.04e-06)		(4.37e-06)
Age of Home		0.0161		0.00774
		(0.0176)		(0.0409)
Size of Property (Acres)		0.894**		6.412**
		(0.361)		(2.977)
% of People with Higher Degree		-8.291***		0.573
		(2.967)		(3.574)
% of People Below Poverty Level		20.19***		23.87***
		(6.798)		(8.544)
Per-capita Income		9.00e-06		-2.52e-05
		(2.58e-05)		(6.65e-05)
% of Renter-Occupied Housing Units		-3.484***		-0.839
		(1.043)		(3.306)
% White		16.06***		11.77***
		(2.680)		(3.175)
Constant	11.46***	0.297	11.75***	-4.412
	(1.015)	(1.836)	(1.632)	(4.302)
Observations	41,011	40,742	21,513	21,513
R-squared	0.145	0.146	0.095	0.099

*In thousands of gallons.

1/ Trimmed sample: Propensity Score estimated using logit model. Observations outside interval [0.06, 0.94] are discarded.

Robust standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

Source: Cobb County Water System, Fulton County Water Service Division, Tax Assessor and 2000 US Census.

Cont.

	(5)	(6)	(7)	(8)
	Matched Sample			
	No Calipers		Calipers	
Social Comparison Treatment	-0.262	0.471	-0.0909	0.431
	(0.314)	(0.309)	(0.284)	(0.330)
Water Use from June - November 2006	-0.378***	-0.401***	-0.360***	-0.380***
	(0.0198)	(0.0191)	(0.0118)	(0.0121)
Water Use in April and May 2007	0.886***	0.869***	0.797***	0.777***
	(0.0712)	(0.0736)	(0.0433)	(0.0435)
Fair Market Value		1.04e-05*		1.51e-05***
		(5.60e-06)		(2.53e-06)
Age of Home		-0.00534		0.0392**
		(0.0248)		(0.0178)
Size of Property (Acres)		0.420		1.360**
		(0.302)		(0.592)
% of People with Higher Degree		-0.819		-3.852
		(2.060)		(2.875)
% of People Below Poverty Level		-7.881		-16.95**
		(5.986)		(8.494)
Per-capita Income		-2.28e-05		-9.45e-05***
		(6.41e-05)		(3.41e-05)
% of Renter-Occupied Housing Units		2.020		3.545*
		(1.639)		(2.155)
% White		6.755***		9.485***
		(1.643)		(3.518)
Constant	3.570***	-2.543	4.041***	-2.195
	(0.809)	(1.881)	(0.441)	(3.542)
Observations	19,971	19,971	14,086	14,086
R-squared	0.284	0.292	0.263	0.272

*In thousands of gallons.

1/ Trimmed sample: Propensity Score estimated using logit model. Observations outside interval [0.06, 0.94] are discarded.

Robust standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

Source: Cobb County Water System, Fulton County Water Service Division, Tax Assessor and 2000 US Census.

Table 22
Non-Experimental Results: OLS Regression with Difference-in-Differences for Technical Information Treatment
Dependent Variable: Summer 2007 – Summer 2006*

	(1)	(2)	(3)	(4)
	Full Sample		Trimmed Sample	
Technical Information Treatment	1.516***	2.565***	2.303***	4.782***
	(0.462)	(0.671)	(0.483)	(0.855)
Water Use from June - November 2006	-0.298***	-0.328***	-0.284***	-0.326***
	(0.0205)	(0.0204)	(0.0309)	(0.0310)
Water Use in April and May 2007	0.113	0.110	0.0983	0.0981
	(0.111)	(0.110)	(0.105)	(0.104)
Fair Market Value		1.76e-05***		2.33e-05***
		(3.54e-06)		(5.11e-06)
Age of Home		0.0413**		0.0696**
		(0.0195)		(0.0275)
Size of Property (Acres)		0.436		4.587***
		(0.312)		(1.324)
% of People with Higher Degree		-4.582		5.509
		(3.004)		(4.010)
% of People Below Poverty Level		22.26***		31.99***
		(6.898)		(8.791)
Per-capita Income		-7.16e-06		-0.000111**
		(2.77e-05)		(5.61e-05)
% of Renter-Occupied Housing Units		-3.565***		-0.994
		(1.059)		(2.442)
% White		15.68***		12.22***
		(2.619)		(2.684)
Constant	10.82***	-3.883*	9.833***	-10.38***
	(1.068)	(2.064)	(1.755)	(3.977)
Observations	41,073	40,801	21,833	21,833
R-squared	0.136	0.137	0.246	0.262

*In thousands of gallons.

1/ Trimmed sample: Propensity Score estimated using logit model. Observations outside interval [0.06, 0.94] are discarded.

Robust standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

Source: Cobb County Water System, Fulton County Water Service Division, Tax Assessor and 2000 US Census.

Cont.

	(5)	(6)	(7)	(8)
	Matched Sample			
	No Calipers		Calipers	
Technical Information Treatment	1.034***	1.838***	0.988***	1.484***
	(0.388)	(0.346)	(0.299)	(0.336)
Water Use from June - November 2006	-0.368***	-0.399***	-0.362***	-0.383***
	(0.0294)	(0.0236)	(0.0132)	(0.0135)
Water Use in April and May 2007	0.986***	0.949***	0.924***	0.904***
	(0.0736)	(0.0776)	(0.0475)	(0.0477)
Fair Market Value		2.16e-05***		1.62e-05***
		(7.23e-06)		(2.57e-06)
Age of Home		0.0420		0.0409**
		(0.0268)		(0.0194)
Size of Property (Acres)		0.204		0.640
		(0.271)		(0.479)
% of People with Higher Degree		-1.867		-6.593**
		(2.074)		(2.884)
% of People Below Poverty Level		-12.77**		-25.39***
		(5.277)		(8.203)
Per-capita Income		-6.52e-05		-3.09e-05
		(6.77e-05)		(4.02e-05)
% of Renter-Occupied Housing Units		0.226		5.160***
		(1.748)		(1.847)
% White		4.739***		12.74***
		(1.571)		(3.703)
Constant	1.522	-3.421**	2.521***	-6.637*
	(0.974)	(1.731)	(0.507)	(3.504)
Observations	20,087	20,087	14,351	14,351
R-squared	0.240	0.256	0.246	0.254

*In thousands of gallons.

1/ Trimmed sample: Propensity Score estimated using logit model. Observations outside interval [0.06, 0.94] are discarded.

Robust standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

Source: Cobb County Water System, Fulton County Water Service Division, Tax Assessor and 2000 US Census.

Table 23
Balance on Covariates using IPW for Social Comparison Treatment

	(1)	(2)	(3)	(4)	(5)	(6)
	Raw Mean		Weighted Mean ATT		Weighted Mean ATE	
	Tr.	Ct.	Tr.	Ct.	Tr.	Ct.
Water Use Jun - Nov 2006	59.6	64.0	59.6	62.5	62.5	63.7
Water Use Apr - May 2007	16.1	15.0	16.1	21.9	16.7	16.7
Fair Market Value	271,704	330,442	271,704	318,608	309,426	327,601
Age of Home	19.4	18.4	19.4	21.7	19.7	19.2
Size of Property (Acres)	0.1	0.5	0.1	0.4	0.2	0.5
% of People with Higher Degree	0.8	0.8	0.8	0.8	0.8	0.8
% of People Below Poverty Level	0.0	0.0	0.0	0.0	0.0	0.0
Per-capita Income	33,276	38,535	33,276	33,922	37,646	37,427
% of Renter-Occupied Housing Units	0.1	0.1	0.1	0.2	0.1	0.1
% White	0.9	0.9	0.9	0.9	0.9	0.9

	(7)	(8)	(9)
	Difference Treatment & Control*		
	Raw	ATT	ATE
Water Use Jun - Nov 2006	4.4	2.9	1.2
Water Use Apr - May 2007	1.1	5.8	0.0
Fair Market Value	58,738	46,903	18,175
Age of Home	1.1	2.3	0.5
Size of Property (Acres)	0.4	0.4	0.3
% of People with Higher Degree	0.1	0.0	0.0
% of People Below Poverty Level	0.0	0.0	0.0
Per-capita Income	5,259	646	219
% of Renter-Occupied Housing Units	0.0	0.1	0.0
% White	0.0	0.0	0.0

	F-test (p-value)		
Water Use Jun - Nov 2006	0.000	0.038	0.177
Water Use Apr - May 2007	0.152	0.440	0.992
Fair Market Value	0.000	0.000	0.000
Age of Home	0.000	0.000	0.030
Size of Property (Acres)	0.000	0.000	0.000
% of People with Higher Degree	0.000	0.013	0.001
% of People Below Poverty Level	0.042	0.000	0.000
Per-capita Income	0.000	0.002	0.325
% of Renter-Occupied Housing Units	0.000	0.000	0.000
% White	0.006	0.000	0.000

* In absolute value

Source: Cobb County Water System, Fulton County Water Service Division, Tax Assessor and 2000 US Census.

Table 24
Balance on Covariates using IPW for Technical Information Treatment

	(1)	(2)	(3)	(4)	(5)	(6)
	Raw Mean		Weighted Mean ATT		Weighted Mean ATE	
	Tr.	Ct.	Tr.	Ct.	Tr.	Ct.
Water Use Jun - Nov 2006	59.4	63.4	59.4	62.0	62.3	63.1
Water Use Apr - May 2007	16.3	14.9	16.3	22.8	16.7	16.7
Fair Market Value	268,871	326,773	268,871	315,683	305,389	324,200
Age of Home	19.5	18.3	19.5	21.4	19.4	19.0
Size of Property (Acres)	0.1	0.4	0.1	0.4	0.2	0.4
% of People with Higher Degree	0.8	0.8	0.8	0.8	0.8	0.8
% of People Below Poverty Level	0.0	0.0	0.0	0.0	0.0	0.0
Per-capita Income	33,226	38,449	33,226	33,761	37,096	37,361
% of Renter-Occupied Housing Units	0.1	0.1	0.1	0.2	0.1	0.1
% White	0.9	0.9	0.9	0.9	0.9	0.9

	(7)	(8)	(9)
	Difference Treatment & Control*		
	Raw	ATT	ATE
Water Use Jun - Nov 2006	4.0	2.6	0.8
Water Use Apr - May 2007	1.4	6.5	0.0
Fair Market Value	57,902	46,812	18,811
Age of Home	1.2	1.9	0.5
Size of Property (Acres)	0.4	0.3	0.3
% of People with Higher Degree	0.1	0.0	0.0
% of People Below Poverty Level	0.0	0.0	0.0
Per-capita Income	5,223	535	265
% of Renter-Occupied Housing Units	0.0	0.1	0.0
% White	0.0	0.0	0.0

	F-test (p-value)		
Water Use Jun - Nov 2006	0.000	0.033	0.342
Water Use Apr - May 2007	0.072	0.444	0.995
Fair Market Value	0.000	0.000	0.000
Age of Home	0.000	0.000	0.029
Size of Property (Acres)	0.000	0.000	0.000
% of People with Higher Degree	0.000	0.005	0.000
% of People Below Poverty Level	0.017	0.000	0.000
Per-capita Income	0.000	0.012	0.232
% of Renter-Occupied Housing Units	0.000	0.000	0.000
% White	0.012	0.000	0.010

* In absolute value

Source: Cobb County Water System, Fulton County Water Service Division, Tax Assessor and 2000 US Census.

Table 25
Non-Experimental Results: IPW for Social Comparison and Technical Information Treatments
Dependent Variable: Summer 2007*

	(1)	(2)	(3)	(4)
	Trimmed Sample			
	ATT		ATE	
	Social Comparison Treatment	Technical Information Treatment	Social Comparison Treatment	Technical Information Treatment
Social Comparison Treatment	-2.941***		-2.266***	
	(1.074)		(0.685)	
Technical Information Treatment		0.119		0.864
		(0.847)		(0.661)
Constant	38.38***	37.36***	39.38***	38.51***
	(1.032)	(0.757)	(0.531)	(0.358)
Observations	21,513	21,833	21,513	21,833
R-squared	0.001	0.000	0.000	0.000

*In thousands of gallons.
Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1
Source: Cobb County Water System, Fulton County Water Service Division, Tax Assessor and 2000 US Census.

Table 26
Non-Experimental Results: Panel Data (Fixed Effect) for Social Comparison and Technical Advice Treatments
Dependent Variable: Monthly Water Use*

	(1)	(2)	(3)	(4)
	Social Comparison Treatment			
	Full	Trimmed	Matched Sample	
	Sample	Sample	No Caliper	Caliper
Social Comparison Treat. * Post Experiment 1/	-0.965***	-0.720***	-0.514***	-0.353**
	(0.0870)	(0.144)	(0.158)	(0.161)
Post Experiment	1.174***	0.973***	0.721***	0.726***
	(0.0770)	(0.136)	(0.153)	(0.155)
Constant	9.020***	8.843***	8.406***	8.421***
	(0.0139)	(0.0223)	(0.0186)	(0.0189)
Observations	697,187	365,721	339,507	239,462
R-squared	0.001	0.000	0.002	0.002
Number of code	41,011	21,513	12,290	8,667

	(5)	(6)	(7)	(8)
	Technical Information Treatment			
	Full	Trimmed	Matched Sample	
	Sample	Sample	No Caliper	Caliper
Technical Information Treat. * Post Experiment 1/	-0.618***	-0.119	-0.109	-0.00257
	(0.0914)	(0.105)	(0.161)	(0.164)
Post Experiment	1.174***	0.827***	0.665***	0.755***
	(0.0770)	(0.0845)	(0.153)	(0.156)
Constant	9.027***	8.817***	8.406***	8.485***
	(0.0140)	(0.0144)	(0.0189)	(0.0193)
Observations	698,240	371,160	341,478	243,966
R-squared	0.001	0.000	0.002	0.003
Number of code	41,073	21,833	12,472	8,918

*In thousands of gallons. Regression analysis period is from May 2006 to September 2007.

1/ Post Experiment represents from June 2007 to September 2007.

2/ Number of observations for matched sample with calipers represent unique households. Repeated observations are taken into account using frequency weights.

Robust standard errors in parentheses.

*** p<0.01, ** p<0.05, * p<0.1

Source: Cobb County Water System, Fulton County Water Service Division.

Table 27
Non-Experimental Bootstrapping Results
Percentage Results within 95% CI Experimental Estimate (*Criterion 2*)*

	(1)	(2)
	Social Comparison Treatment	Technical Information Treatment
Full Sample		
- OLS & Full Sample	24%	22%
- Panel Data & Full Sample	0%	0%
Trimmed Sample		
- OLS & Trimming Rule (Logit)	3%	2%
- OLS & Trimming Rule (Probit)	17%	19%
- OLS & Trimming Rule & IPW (ATT, Logit)	15%	32%
- OLS & Trimming Rule & IPW (ATE, Logit)	29%	18%
- OLS & Trimming Rule & IPW (ATT, Probit)	2%	1%
- OLS & Trimming Rule & IPW (ATE, Probit)	45%	28%
- Panel Data & Trimming Rule (Logit)	2%	41%
- Panel Data & Trimming Rule (Probit)	0%	2%
Matching Sample		
- OLS & Matching without Calipers	60%	14%
- OLS & Matching with Calipers	48%	70%
- Panel Data & Matching without Calipers	16%	50%
- Panel Data & Matching with Calipers	73%	79%

* Results based on 500 repetitions.

Figure 4
Optimal Trimming Rule: Observations Within and Outside Range

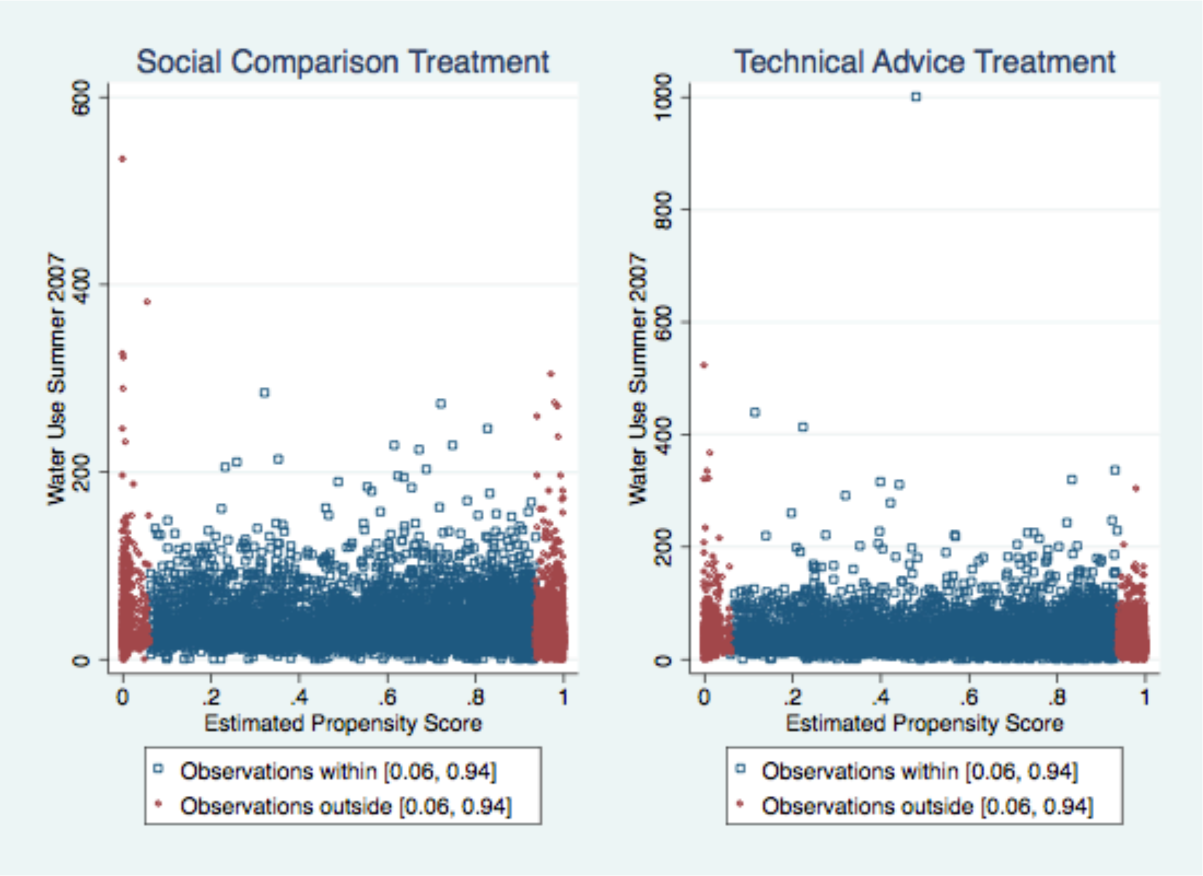


Figure 5
Matching Rule: Observations Discarded using Covariate Matching with and without Calipers

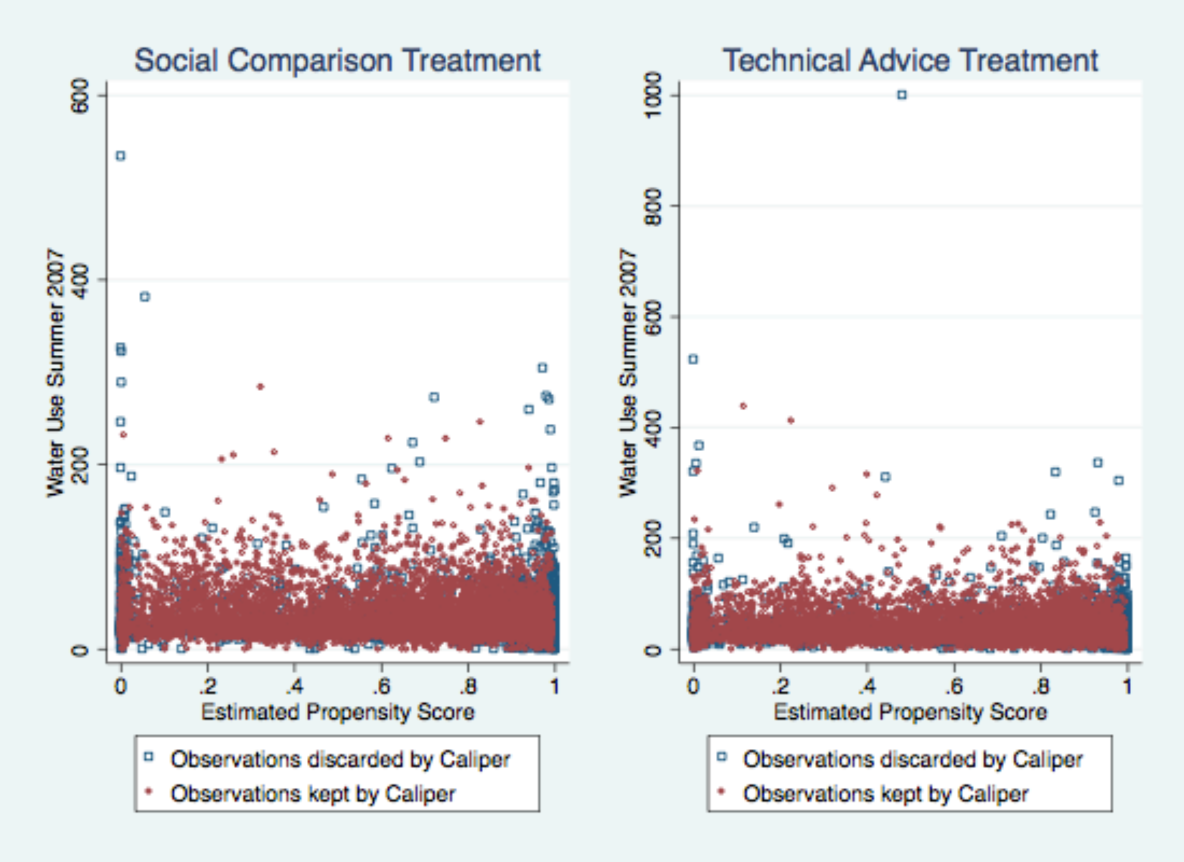


Figure 6
Estimated Propensity Score Histogram for Social Comparison and Technical Information Treatments

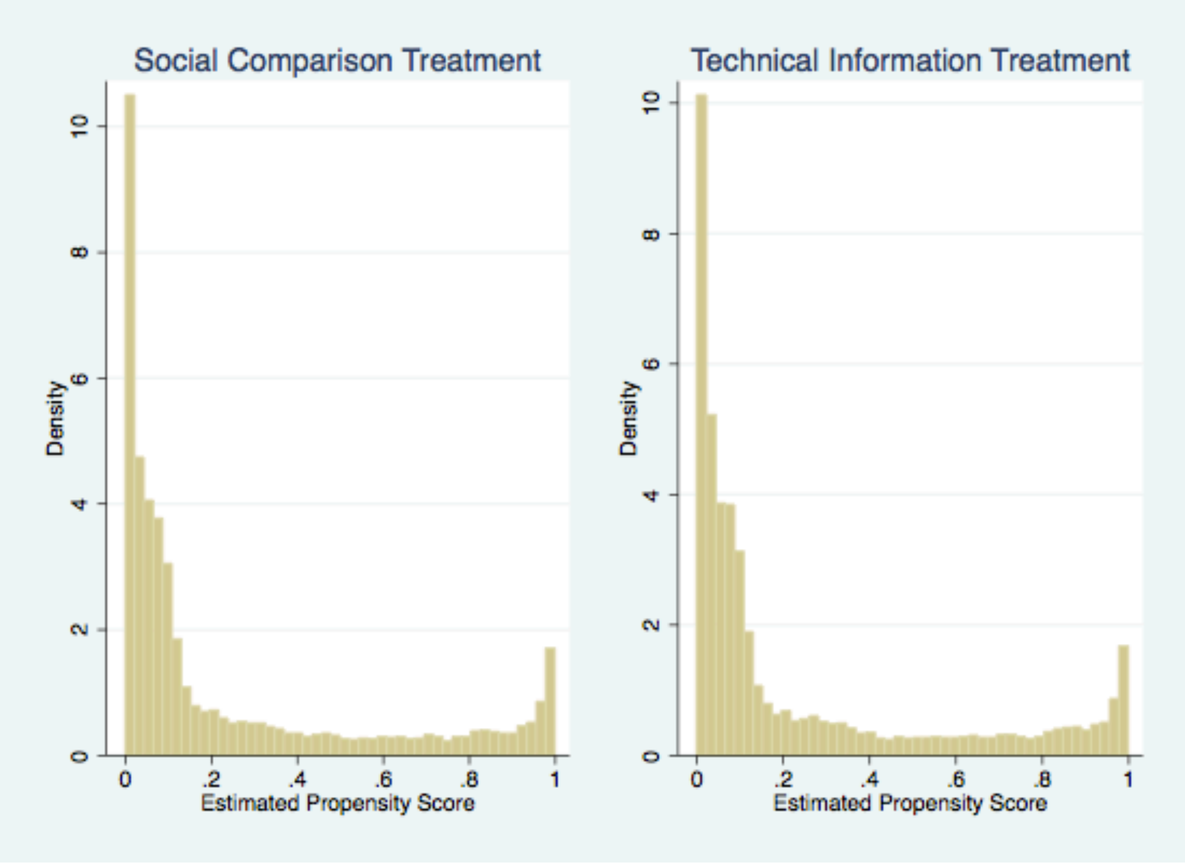
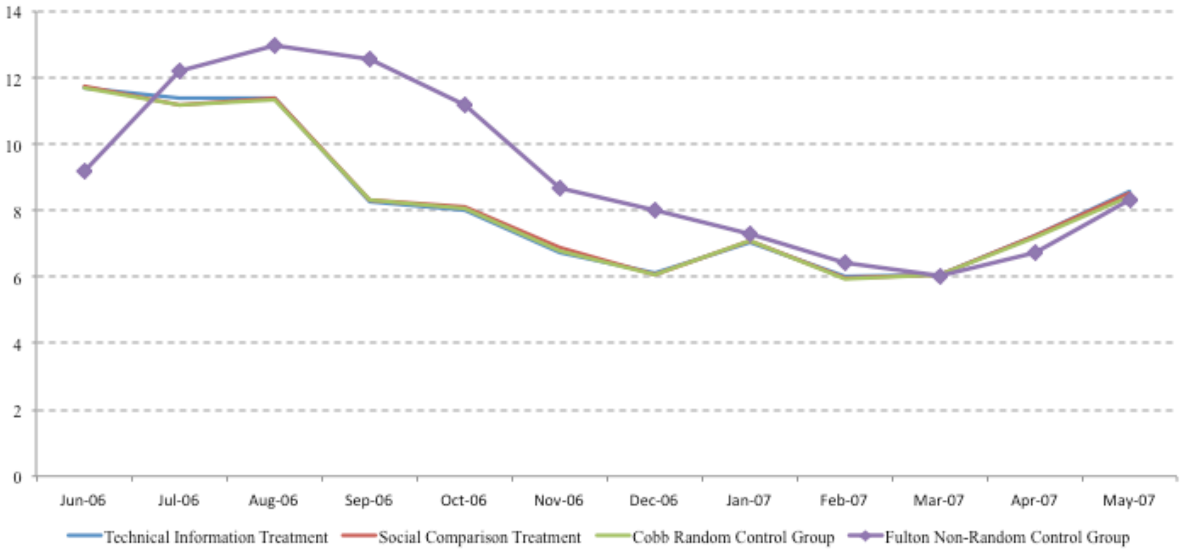
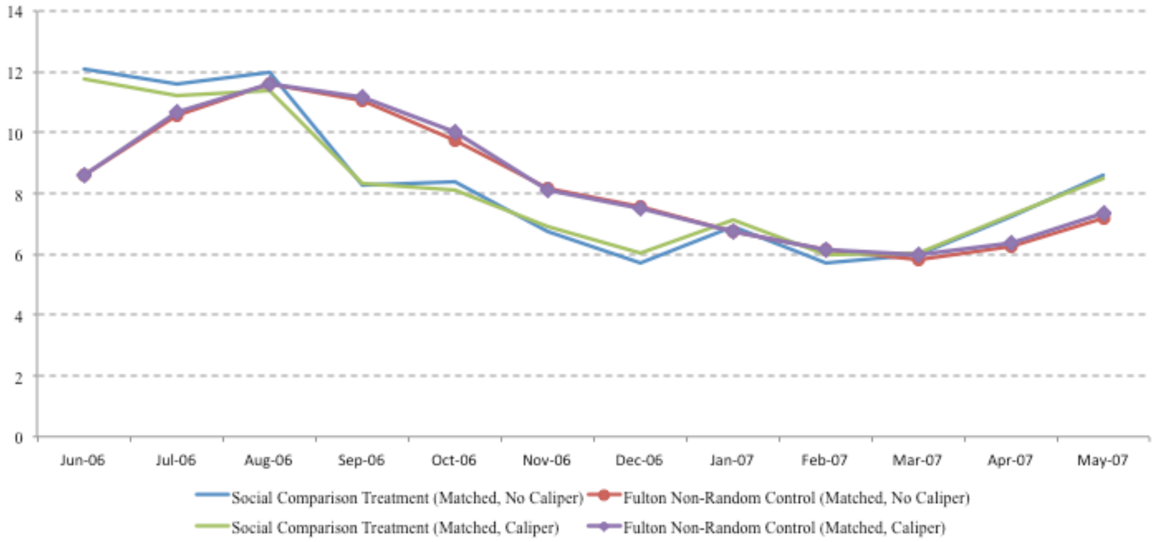


Figure 7
Pre-Treatment Mean Water Consumption
Social Comparison Treatment, Technical Information Treatment, Cobb Random Control Group, and Fulton non-Random Control Group*



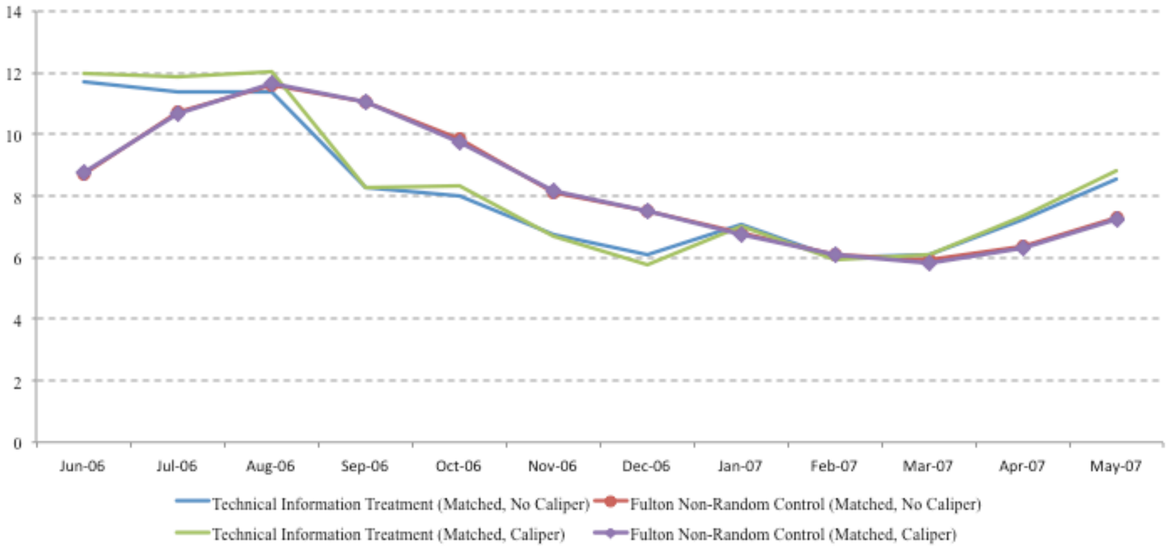
* In thousands of gallons.

Figure 8
Pre-Treatment Mean Water Consumption: Social Comparison Treatment and Fulton non-Random Control Group
Matched Sample with and without Calipers*



* In thousands of gallons.

Figure 9
Pre-Treatment Mean Water Consumption: Technical Information Treatment and Fulton non-Random Control Group
Matched Sample with and without Calipers*



* In thousands of gallons.

Figure 10
Non-Experimental Bootstrapping of Treatment Effect:
OLS for Full Sample and Matched Sample

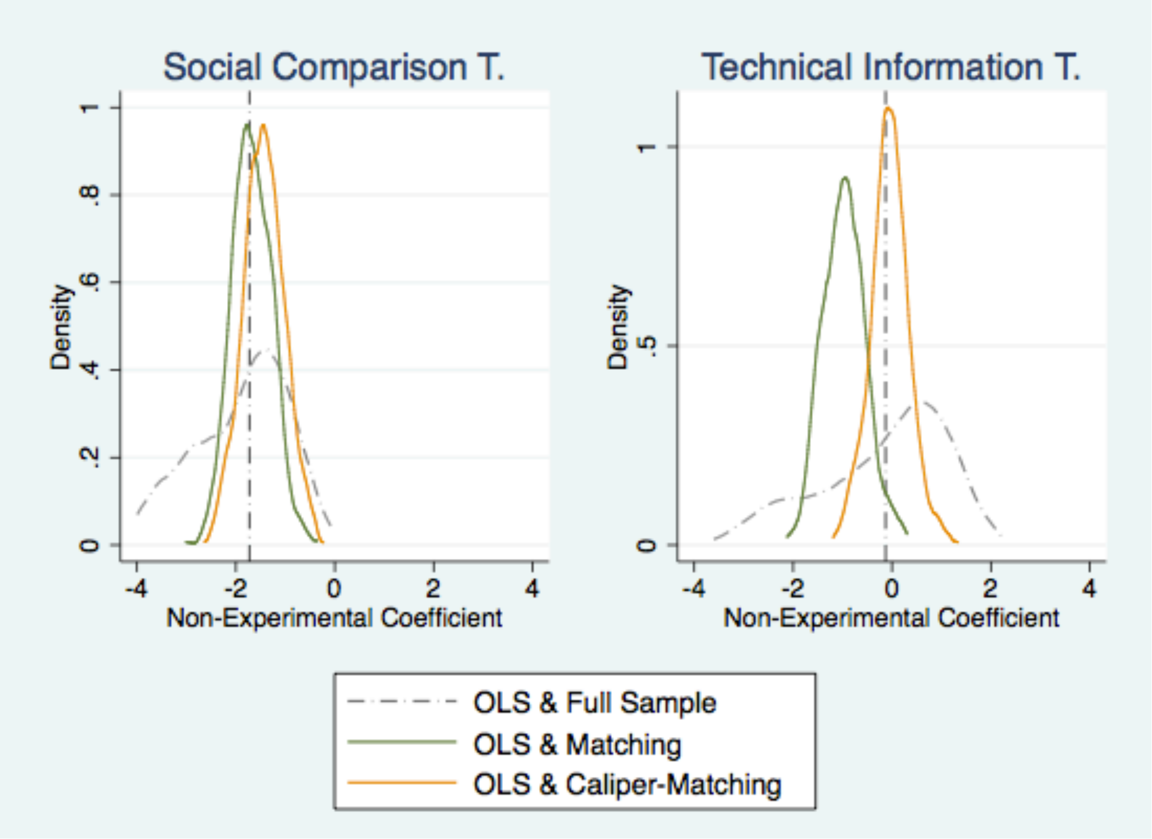


Figure 11
Non-Experimental Bootstrapping of Treatment Effect:
OLS and IPW for Full Sample and Trimmed Sample

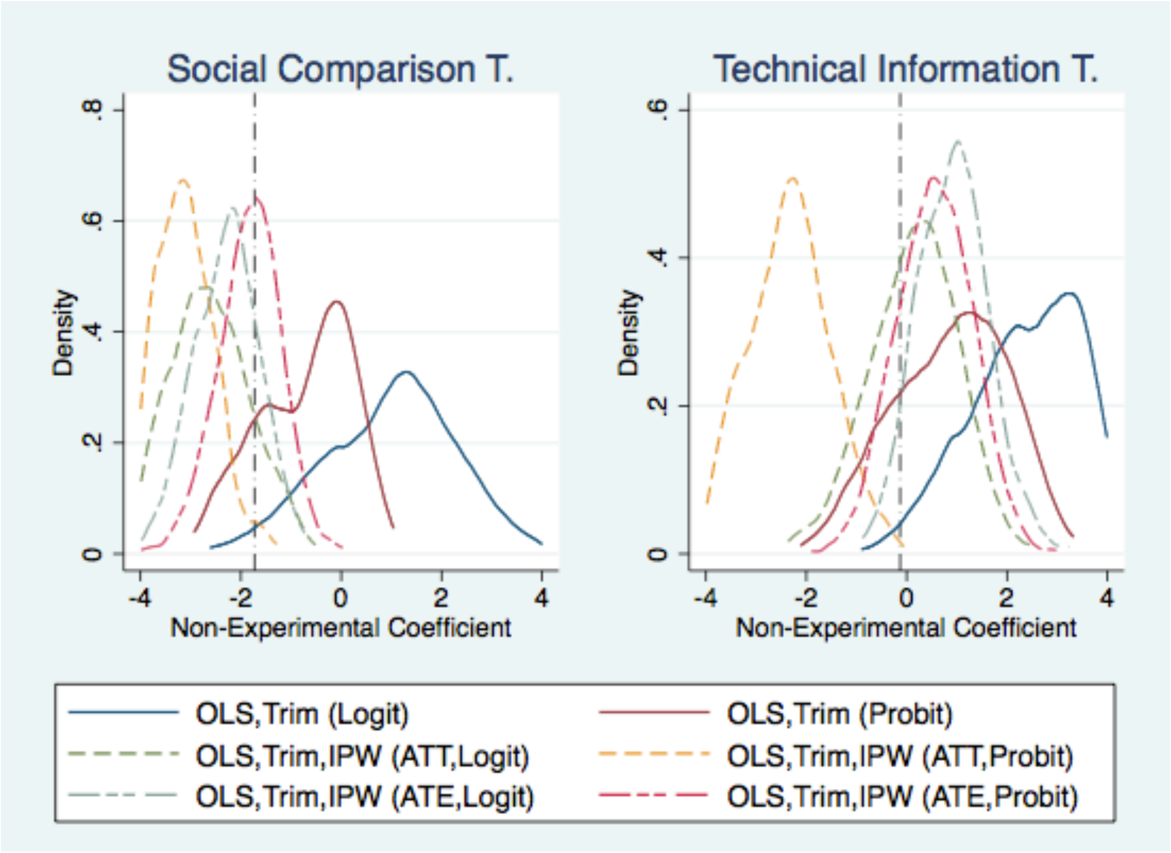
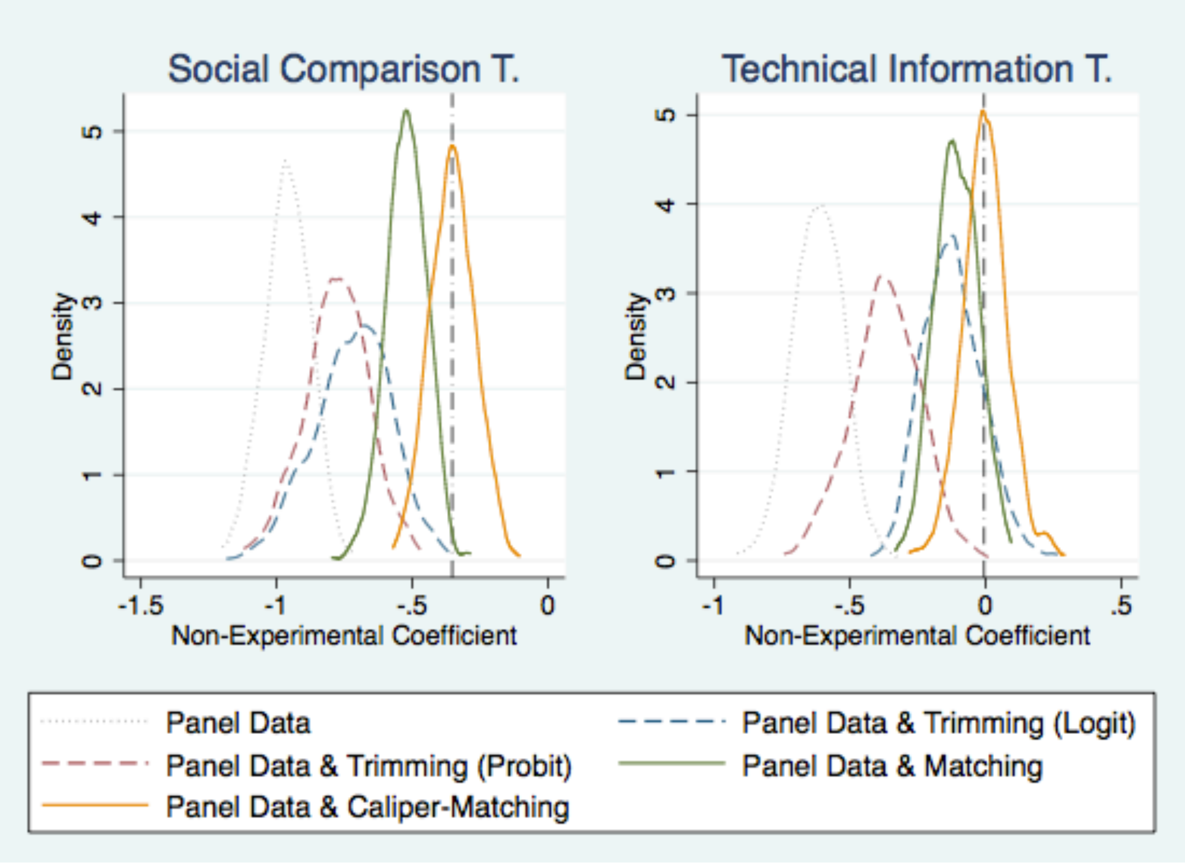


Figure 12
Non-Experimental Bootstrapping of Treatment Effect:
Panel Data for Full, Trimmed and Matched Sample



Appendix 10
Covariate Matching Single Difference Estimates using Different Subset of Covariates

	(1)	(2)	(3)	(4)
	Social Comparison Treatment		Technical Information Treatment	
	No Caliper	Caliper	No Caliper	Caliper
1. Previous Water Use variables 1/	-4.775	-4.748	-3.156	-3.156
Standard Error	(0.262)	(0.261)	(0.291)	(0.291)
1.1. With Bias Correction	-4.814	-4.778	-3.187	-3.187
Standard Error	(0.347)	(0.345)	(0.376)	(0.376)
2. Property variables 2/	4.853	4.917	7.075	7.029
Standard Error	(0.369)	(0.366)	(0.374)	(0.365)
2.1. With Bias Correction	5.493	5.760	7.676	8.029
Standard Error	(1.320)	(1.319)	(1.183)	(1.180)
3. Neighborhood variables 3/	-4.293	-2.162	-2.840	-0.237
Standard Error	(0.409)	(0.486)	(0.438)	(0.527)
3.1. With Bias Correction	-2.892	-1.831	-1.382	0.143
Standard Error	(1.159)	(0.861)	(1.164)	(0.873)
4. Property and Neighborhood variables	-0.013	2.163	1.961	4.375
Standard Error	(0.365)	(0.413)	(0.369)	(0.439)
4.1. With Bias Correction	4.953	5.770	7.073	9.116
Standard Error	(1.258)	(1.101)	(1.261)	(1.133)

1/ Aggregate water use from May 2006 to October 2006, Aggregate water use for March and April 2007.

2/ Fair market value of home ($), Property size (acres), Age of home (years)

3/ Per-capita income ($), Percent of adults over 25 years old with college education or higher, Percent of people living below poverty line, Percent of population that is white, and Percent of renter-occupied housing units.

Appendix 11
Non-Experimental Result: Single Difference with OLS Regression for Social Comparison Treatment
Dependent Variable: Summer 2007*

	(1)	(2)	(3)	(4)	(5)	(6)
	Full Sample			Trimmed Sample		
Social	0.598	-2.660***	-1.704**	6.268***	-0.444	5.936***
Comparison Treat.	(0.853)	(0.667)	(0.707)	(0.757)	(0.810)	(1.075)
Fair Market Value	6.2e-05***		6.1e-05***	6.3e-05***		6.0e-05***
	(9.47e-06)		(1.07e-05)	(3.41e-06)		(4.69e-06)
Age of Home	-0.00564		-0.0680	0.00441		-0.0677
	(0.0603)		(0.0600)	(0.0342)		(0.0467)
Size of Property	2.918***		2.236***	18.34***		16.49***
(Acres)	(0.820)		(0.742)	(2.868)		(3.070)
% of People with		-32.97***	-38.85***		-54.08***	-27.46***
Higher Degree		(3.505)	(3.484)		(8.175)	(5.312)
% of People		28.57***	12.19		-44.82***	-20.98**
Below Poverty		(8.779)	(9.033)		(12.39)	(10.29)
Per-capita Income		0.000678***	0.000152*		0.00125***	0.000315***
		(2.58e-05)	(8.72e-05)		(6.10e-05)	(7.30e-05)
% of Renter-		-1.245	-1.963		9.477***	12.90***
Occupied		(1.317)	(1.315)		(2.903)	(3.636)
% White		33.18***	38.74***		12.32***	28.32***
		(2.981)	(2.960)		(3.800)	(3.365)
Constant	17.71***	11.43***	12.49***	9.934***	26.56***	-1.244
	(4.450)	(2.323)	(2.386)	(1.322)	(4.923)	(5.271)
Observations	40,742	41,011	40,742	21,513	21,513	21,513
R-squared	0.092	0.027	0.099	0.066	0.022	0.069

*In thousands of gallons.

1/ Trimmed sample: Propensity Score estimated using logit model. Observations outside interval [0.06, 0.94] are discarded.

Robust standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

Source: Cobb County Water System, Fulton County Water Service Division, Tax Assessor and 2000 US Census.

Cont.

	(7)	(8)	(9)	(10)	(11)	(12)
	Matched Sample					
	No Calipers			Calipers		
Social Comparison Treat.	1.683***	2.660***	1.745***	1.807***	1.628***	1.438***
	(0.429)	(0.406)	(0.419)	(0.422)	(0.411)	(0.435)
Fair Market Value	7.8e-05***		7.7e-05***	8.1e-05***		8.4e-05***
	(5.77e-06)		(7.79e-06)	(2.87e-06)		(3.74e-06)
Age of Home	0.0548**		0.0592*	0.128***		0.148***
	(0.0257)		(0.0348)	(0.0213)		(0.0250)
Size of Property (Acres)	0.251		0.213	1.079		0.717
	(0.526)		(0.537)	(0.759)		(0.777)
% of People with Higher Degree		-2.655	-8.213***		23.65***	-1.572
		(3.225)	(2.983)		(3.876)	(3.912)
% of People Below Poverty		7.676	-27.08***		36.38***	-30.76***
		(7.323)	(8.173)		(11.63)	(11.48)
Per-capita Income		0.000836***	6.09e-05		0.000475***	-0.00014***
		(6.34e-05)	(8.72e-05)		(5.51e-05)	(4.82e-05)
% of Renter-Occupied		8.184***	-0.0415		6.335**	0.0267
		(1.893)	(2.140)		(2.923)	(2.820)
% White		11.00***	3.775*		43.22***	16.89***
		(2.003)	(2.136)		(4.754)	(4.888)
Constant	11.77***	-2.483	13.64***	9.097***	-41.03***	-0.293
	(2.105)	(2.135)	(2.392)	(0.999)	(4.582)	(4.855)
Observations	19,971	19,971	19,971	14,086	14,086	14,086
R-squared	0.226	0.066	0.228	0.198	0.060	0.200

*In thousands of gallons.

1/ Trimmed sample: Propensity Score estimated using logit model. Observations outside interval [0.06, 0.94] are discarded.

Robust standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

Source: Cobb County Water System, Fulton County Water Service Division, Tax Assessor and 2000 US Census.

Appendix 12
Non-Experimental Result: Single Difference with OLS Regression for Technical Information Treatment
Dependent Variable: Summer 2007*

	(1)	(2)	(3)	(4)	(5)	(6)
		Full Sample			Trimmed Sample	
Technical	2.379**	-0.552	0.291	9.321***	3.235***	9.537***
Information Treat.	(0.941)	(0.705)	(0.708)	(0.816)	(0.648)	(0.906)
Fair Market Value	6.6e-05***		6.5e-05***	7.3e-05***		7.0e-05***
	(1.04e-05)		(1.18e-05)	(4.19e-06)		(5.03e-06)
Age of Home	0.0249		-0.0344	0.0804***		0.00712
	(0.0657)		(0.0657)	(0.0299)		(0.0352)
Size of Property	2.454***		1.822***	18.16***		17.16***
(Acres)	(0.748)		(0.671)	(1.531)		(1.646)
% of People with		-29.95***	-36.03***		-43.35***	-18.16***
Higher Degree		(3.538)	(3.457)		(5.388)	(5.524)
% of People		25.60***	9.343		-45.11***	-20.67**
Below Poverty		(9.025)	(9.161)		(10.78)	(10.30)
Per-capita Income		0.000694***	0.000137		0.00129***	0.000245***
		(2.65e-05)	(9.62e-05)		(6.43e-05)	(6.28e-05)
% of Renter-		-0.352	-1.804		15.49***	15.92***
Occupied		(1.333)	(1.359)		(2.890)	(2.899)
% White		32.02***	38.55***		7.818***	24.84***
		(2.956)	(2.936)		(2.894)	(2.969)
Constant	15.98***	9.133***	9.153***	5.090***	18.68***	-9.029*
	(4.862)	(2.557)	(2.649)	(1.969)	(4.585)	(4.882)
Observations	40,801	41,073	40,801	21,833	21,833	21,833
R-squared	0.093	0.026	0.099	0.177	0.049	0.181

*In thousands of gallons.
1/ Trimmed sample: Propensity Score estimated using logit model. Observations outside interval [0.06, 0.94] are discarded.
Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1
Source: Cobb County Water System, Fulton County Water Service Division, Tax Assessor and 2000 US Census.

Cont.

	(7)	(8)	(9)	(10)	(11)	(12)
	Matched Sample					
	No Calipers			Calipers		
Technical Information Treat.	4.017***	4.645***	3.799***	3.673***	3.607***	3.266***
	(0.466)	(0.458)	(0.407)	(0.455)	(0.458)	(0.473)
Fair Market Value	8.8e-05***		8.9e-05***	9.1e-05***		9.5e-05***
	(5.37e-06)		(7.20e-06)	(2.82e-06)		(3.56e-06)
Age of Home	0.0958***		0.115***	0.154***		0.179***
	(0.0236)		(0.0302)	(0.0226)		(0.0269)
Size of Property (Acres)	0.577		0.474	0.767		0.224
	(0.482)		(0.482)	(0.769)		(0.776)
% of People with Higher Degree		-3.333	-8.651***		22.34***	-4.262
		(3.023)	(2.802)		(4.127)	(4.061)
% of People Below Poverty		-5.447	-38.83***		25.07**	-51.70***
		(7.857)	(7.598)		(12.28)	(11.80)
Per-capita Income		0.000886***	-1.72e-05		0.000574***	-0.00015***
		(5.95e-05)	(7.09e-05)		(6.35e-05)	(5.54e-05)
% of Renter-Occupied		9.961***	-1.641		11.07***	5.435**
		(1.837)	(2.116)		(2.888)	(2.741)
% White		9.692***	3.626*		45.38***	19.60***
		(2.023)	(2.064)		(4.975)	(4.944)
Constant	7.812***	-2.722	12.56***	6.247***	-45.47***	-3.483
	(1.948)	(2.249)	(2.193)	(1.011)	(4.731)	(4.759)
Observations	20,087	20,087	20,087	14,351	14,351	14,351
R-squared	0.220	0.066	0.222	0.195	0.063	0.198

*In thousands of gallons.

1/ Trimmed sample: Propensity Score estimated using logit model. Observations outside interval [0.06, 0.94] are discarded.

Robust standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

Source: Cobb County Water System, Fulton County Water Service Division, Tax Assessor and 2000 US Census.

Appendix 13
Non-Experimental Result: Inverse Probability Weighting Regression (IPW) using Probit Models for Social Comparison and Technical Information Treatments
Dependent Variable: Summer 2007*

	(1)	(2)	(3)	(4)
	Trimmed Sample			
	ATT		ATE	
	Social Comparison Treatment	Technical Information Treatment	Social Comparison Treatment	Technical Information Treatment
Social Comparison Treatment	-3.434***		-1.779***	
	(0.829)		(0.643)	
Technical Information Treatment		-2.375***		0.552
		(0.895)		(0.733)
Constant	38.85***	39.35***	39.78***	39.80***
	(0.775)	(0.822)	(0.471)	(0.469)
Observations	25,150	25,945	25,150	25,945
R-squared	0.001	0.000	0.000	0.000

*In thousands of gallons.
Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1
Source: Cobb County Water System, Fulton County Water Service Division, Tax Assessor and 2000 US Census.

Appendix 14
Summary of Non-Experimental Results using Summer 2006 as Previous Water Use

	Social Comparison Treatment			
	Full Sample	Trimmed Sample	Matching Sample	Caliper-Matching Sample
Single Difference Estimator				
A) Single Difference	-7.026***	-3.338***	-0.220	1.289***
B) Single Difference with Covariate Matching			3.856***	3.658***
C) Single Difference with IPW				
C.1) ATT weights		-1.666*	5.889***	6.575***
C.2) ATE weights		-1.026	5.837***	7.017***
D) Single Difference with OLS				
D.1) Only Previous Water Use	-0.939***	-0.704	1.623***	2.060***
D.2) Previous Water Use, Property and Neighborhood Variables	0.0352	2.101**	2.750***	2.833***

Difference-in-Differences Estimator				
E) Difference-in-Differences	2.887***	0.783	2.553***	2.457***
F) Difference-in-Differences with Covariate Matching			3.856***	3.658***
G) Difference-in-Differences with IPW				
G.1) ATT weights		0.506	0.946**	0.993**
G.2) ATE weights		-0.259	0.388	0.460
H) Difference-in-Difference with OLS				
H.1) Only Previous Water Use	-0.939***	-0.704	1.623***	2.060***
H.2) Previous Water Use, Property and Neighborhood Variables	0.0352	2.101**	2.750***	2.833***
H.3) Only Property and Neighborhood Variables (No Previous Water Use)	1.340**	0.391	2.355***	2.446***
I) Difference-in-Differences with Fixed-Effects Panel Data	-0.965***	-0.760***	-0.447***	-0.293*

Notes:

- Panel A), B), C), E), F), G) are comparable to Table 9 (Social Comparison Treatment) and Table 10 (Technical Information Treatment).
- Panel D) is comparable to Table 11 (Social Comparison Treatment) and Table 12 (Technical Information Treatment).
- Panel H) is comparable to Table 13 (Social Comparison Treatment) and Table 14 (Technical Information Treatment).
- Panel I) is comparable to Table 18.

Cont.

	Technical Information Treatment			
	Full Sample	Trimmed Sample	Matching Sample	Caliper-Matching Sample
Single Difference Estimator				
A) Single Difference	-5.519***	-0.435	1.233***	2.616***
B) Single Difference with Covariate Matching			5.573***	5.214***
C) Single Difference with IPW				
C.1) ATT weights		1.593**	7.777***	7.994***
C.2) ATE weights		2.345***	8.072***	8.382***
D) Single Difference with OLS				
D.1) Only Previous Water Use	0.663*	1.894***	3.238***	3.527***
D.2) Previous Water Use, Property and Neighborhood Variables	2.001***	4.479***	4.220***	4.304***

Difference-in-Differences Estimator				
E) Difference-in-Differences	4.383***	3.292***	4.109***	3.940***
F) Difference-in-Differences with Covariate Matching			5.573***	5.214***
G) Difference-in-Differences with IPW				
G.1) ATT weights		2.968***	3.160***	2.931***
G.2) ATE weights		2.663***	2.983***	2.612***
H) Difference-in-Difference with OLS				
H.1) Only Previous Water Use	0.663*	1.894***	3.238***	3.527***
H.2) Previous Water Use, Property and Neighborhood Variables	2.001***	4.479***	4.220***	4.304***
H.3) Only Property and Neighborhood Variables (No Previous Water Use)	3.248***	1.847**	3.785***	3.995***
I) Difference-in-Differences with Fixed-Effects Panel Data	-0.618***	-0.147	-0.0670	0.0178

Notes:

- Panel A), B), C), E), F), G) are comparable to Table 9 (Social Comparison Treatment) and Table 10 (Technical Information Treatment).
- Panel D) is comparable to Table 11 (Social Comparison Treatment) and Table 12 (Technical Information Treatment).
- Panel H) is comparable to Table 13 (Social Comparison Treatment) and Table 14 (Technical Information Treatment).
- Panel I) is comparable to Table 18.

Appendix 15
Non-Experimental Bootstrapping Results
Percentage Results within 95% CI Experimental Estimate *and* Rejection Rates*

	(1)	(2)
	Social Comparison Treatment (Within Experimental 95% CI & Null Rejection)	Technical Information Treatment (Within Experimental 95% CI & No Null Rejection)
Full Sample		
- OLS & Full Sample	23%	22%
- Panel Data & Full Sample	0%	0%
Trimmed Sample		
- OLS & Trimming Rule (ATT, Logit)	2%	2%
- OLS & Trimming Rule (Probit)	17%	19%
- OLS & Trimming Rule & IPW (ATT, Logit)	12%	32%
- OLS & Trimming Rule & IPW (ATE, Logit)	29%	18%
- OLS & Trimming Rule & IPW (ATT, Probit)	2%	1%
- OLS & Trimming Rule & IPW (ATE, Probit)	44%	28%
- Panel Data & Trimming Rule (Logit)	2%	41%
- Panel Data & Trimming Rule (Probit)	0%	2%
Matching Sample		
- OLS & Matching without Calipers	60%	14%
- OLS & Matching with Calipers	48%	70%
- Panel Data & Matching without Calipers	16%	50%
- Panel Data & Matching with Calipers	73%	79%

* Results based on 500 repetitions.

VITA

Juan José Miranda Montero is originally from Peru. He was born on March 8, 1978 in Chimbote, Ancash, Peru. He holds a Bachelor of Arts in Economics from Universidad del Pacifico in Lima, Peru and a Master of Arts in Economics from Georgia State University.

Juan José began Georgia State's Doctoral program in 2007 to study Environmental Economics and Experimental Economics. He has worked as a graduate research assistant for Dr. Paul J. Ferraro. During his doctoral program, Juan José received a University Fellowship, and he has also worked as a consultant for the World Bank and the Global Environmental Facility.

Prior to joining the doctoral program, Juan José has worked as assistant researcher and research associate at the Instituto de Estudios Peruanos (IEP) on topics of environmental economics, development economics, and regulatory economics.

Juan José received his Doctor of Philosophy degree in Economics from Georgia State University in July 2012. He has since accepted the position of consultant with the Office of Strategic Planning and Development Effectiveness of the Inter-American Development Bank in Washington, D.C.