Economics Letters 151 (2017) 82–90

Contents lists available at ScienceDirect

Economics Letters

journal homepage: www.elsevier.com/locate/ecolet

A cautionary tale on using panel data estimators to measure program impacts

Casey J. Wichman [a,*], Paul J. Ferraro [b]

[a] Resources for the Future, 1616 P St. NW, Washington, DC 20036, United States

[b] Bloomberg School of Public Health, Carey Business School, Whiting School of Engineering, Johns Hopkins University, 100 International St., Baltimore, MD 21202, United States

HIGHLIGHTS

- We compare experimental and observational estimates of environmental program impact.
- We expand the sample of comparison units to improve covariate balance.
- Despite similarity of covariates and baseline trends, bias of the estimator worsens.
- Fixed-effects panel estimators and indirect tests of their validity are no panacea.

ARTICLE INFO

Article history:
Received 31 May 2016
Received in revised form
19 November 2016
Accepted 22 November 2016
Available online 13 December 2016

JEL classification:
C52
C93
D12
H42
Q25

Keywords:
Design replication
Program evaluation
Matching
Panel data
Water conservation

ABSTRACT

We compare experimental and nonexperimental estimates from a social and informational messaging experiment. Our results show that applying a fixed effects estimator in conjunction with matching to pre-process nonexperimental comparison groups cannot replicate an experimental benchmark, despite parallel pre-intervention trends and good covariate balance. The results are a stark reminder about the role of untestable assumptions – in our case, conditional bias stability – in drawing causal inferences from observational data, and the dangers of relying on single studies to justify program scaling-up or canceling.

1. Introduction

Researchers using observational data often confront the question: what is the ideal experiment to identify my causal relationship? Less common is the question: how accurate is the estimate of my observational design relative to an experimental benchmark? To consider this question, researchers use "design replications", or "within-study designs", in which causal estimates from randomized experiments are compared to estimates from nonexperimental replications (Cook et al., 2008). In theory, nonexperimental designs can perform as well as experimental designs. These design replications allow researchers to examine the validity of the assumptions used to identify causal effects in specific nonexperimental contexts. How best to interpret the results of design replications has, however, been contentious (Lalonde, 1986; Heckman et al., 1997; Smith and Todd, 2005; Dehejia, 2005).

One source of contention is the failure of design replication studies to consider the sensitivity of their results to the choice of sample (Smith and Todd, 2005).[1] In a design replication study

* Corresponding author.
E-mail addresses: wichman@rff.org (C.J. Wichman), pferraro@jhu.edu (P.J. Ferraro).

http://dx.doi.org/10.1016/j.econlet.2016.11.029
0165-1765/

[1] Or, as a referee pointed out, one might interpret Smith and Todd's (2005) analysis as changing the *population*, rather than the sample. We explore this

using a fixed effects panel data (FEPD) estimator in conjunction with matching to pre-process the comparison group data, Ferraro and Miranda (forthcoming) show that an observational design using comparison households from a neighboring county can replicate results from an experimental design.[2] Through a bootstrapping exercise, they further demonstrate that the treatment effect estimates are not sensitive to the choice of sample within the two counties.

An alternative way to assess sensitivity to sample choice is to expand the pool of untreated units. Conventional wisdom suggests that increasing the number of comparison units from which to select a comparison group should (weakly) improve nonexperimental designs (Heckman et al., 1997). We assess this wisdom by extending the design of Ferraro and Miranda with the addition of a second group of untreated households, which are observationally more similar to the treated households. Including additional comparison households greatly improves covariate balance and yields parallel pre-treatment trends in outcomes. Despite these improvements, however, we find that the FEPD estimator, with or without pre-processing the data, performs worse: it no longer replicates the experimental benchmark.

2. An experimental benchmark and nonexperimental comparison groups

Our experimental benchmark comes from a randomized controlled trial (RCT) with over 100,000 households in Cobb County, Georgia (Ferraro and Price, 2013). In the RCT, a water utility sent messages to households to induce voluntary reductions in water use. Each treatment group comprised approximately 11,700 households and the control group, 71,600 households. Treatment assignment was randomized at the household level within nearly 400 meter route strata (i.e., small neighborhoods).[3]

We examine two of Ferraro and Price's treatments: (i) a **technical information treatment**, which instructed households on strategies to reduce water use; and (ii) a **social comparison treatment**, which augmented the technical information with social norm-based encouragement and a social comparison in which own consumption was compared to median county consumption. In the original experiment, the social comparison treatment induced a large (approximately 5%) statistically significant reduction in water consumption while the technical information treatment displayed a small (approximately 0.5%) statistically insignificant effect.

To construct nonexperimental comparison groups, we use households from neighboring Fulton County (used by Ferraro and Miranda, forthcoming), and nearby Gwinnett County. Cobb, Fulton, and Gwinnett counties had similar water pricing policies and the same water sources, weather patterns, state and metro regulatory environments, and other regional confounding factors during the experiment. To our knowledge, there were no contemporaneous policy changes in the comparison counties. We believe these comparison groups thus meet the Heckman et al. and Cook et al. criteria for effective observational designs.

3. Empirical strategy

Our identification strategy uses repeated observations on households to control for unobserved and unchanging characteristics that are related to water consumption and exposure to the treatment (Angrist and Krueger, 1999). Our design relies on the common linear, additive FEPD estimator,

$$w_{it} = \alpha + \boldsymbol{A}_i'\gamma + \boldsymbol{X}_{it}'\beta + \delta Treat_{it} + \lambda_{\boldsymbol{t}} + \varepsilon_{it}, \quad (1)$$

where w_{it} is monthly water use for household i at time t; $\boldsymbol{A}_i$ is a vector of fixed (time-invariant) household characteristics; $\boldsymbol{X}_{\boldsymbol{it}}$ is a vector of time-varying household characteristics; $Treat_{it}$ is a treatment indicator; and $\lambda_{\boldsymbol{t}}$ are time fixed effects. Under an assumption of conditional bias stability, Eq. (1) provides an unbiased estimator of the Average Treatment Effect, δ, which was also the estimand estimated by the RCT. Conditional bias stability asserts that conditional on $\boldsymbol{X}_{\boldsymbol{it}}$, pre-program differences in outcomes between treatment and comparison groups are stable across post-program periods. Ferraro and Miranda make the case for the plausibility of this assumption in the study context.

3.1. Data and samples

We use household water consumption data from the Cobb County Water System, the Fulton County Water Service Division, and the Gwinnett Department of Water Resources. We have thirteen months of pre-treatment data (May 2006–May 2007) and four months of post-treatment data (June–September 2007). The county tax assessor databases provide home and property characteristics, and the 2000 US Census provides data on neighborhood characteristics at the block-group level.

Table 1 shows average water consumption in thousands of gallons during key watering seasons for Cobb households in the experiment, and for Fulton and Gwinnett households. We also consider covariates that are observable to policymakers and that theory or empirical studies suggest could be important confounders in a study on water conservation (e.g., Ferraro and Miranda, 2014; Wichman et al., 2016). Overall, Gwinnett households appear to be more similar to treatment households along water use and socioeconomic characteristics than do Fulton households.

4. Observational measuring sticks

Drawing causal inferences in any nonexperimental design requires making untestable assumptions (e.g., model dependence, unconfoundedness, and so on).[4] To overcome model dependence, researchers are increasingly using matching techniques to reweight the sample so that treatment and comparison groups are similar and, thus, rely less heavily on parametric assumptions (Ho et al., 2007). Furthermore, observing parallel trends in outcomes prior to treatment is commonly used to support the conditional bias stability assumption. As in Ferraro and Miranda, we focus on these two empirical heuristics in our analysis.

4.1. Does trimming and matching improve covariate balance?

Following Ferraro and Miranda (forthcoming), we first use the **full sample** of treated and comparison households. Second, we construct a **trimmed sample** using the optimal trimming rule of Crump et al. (2009) to remove observations with extreme propensity scores.[5] Third, we construct two **matched samples**. We use nearest-neighbor (1:1) Mahalanobis covariate matching

point empirically by examining two comparison groups separately (i.e., as distinct populations) as well as jointly (i.e., as different draws from the same population).

[2] Pre-processing in our context refers to matching or trimming to reweight the sample prior to applying a parametric estimator.

[3] For more details on the experiment and randomization, see Ferraro and Price (2013).

[4] Causal inference in experimental designs also relies on untestable assumptions (Heckman and Smith, 1995), but fewer than are required in nonexperimental designs.

[5] Based on a logit model, our optimal trimming rule discards observations with estimated propensity scores outside the interval [0.03, 0.97].

Table 1
Descriptive statistics for Cobb, Gwinnett, and Fulton counties.

	(1) Cobb County	(2)	(3)	(4) Gwinnett Co.	(5) Fulton Co.
	Technical information treatment	Social comparison treatment	Experimental control	Non-experimental comparison	Non-experimental comparison
Water use (consumption) variables					
May–Oct 2006	58.32	58.45	58.24	55.83	67.24
	(39.77)	(40.80)	(41.13)	(38.61)	(55.68)
Mar–May 2007	27.45	27.01	27.73	25.23	24.77
	(19.89)	(19.06)	(79.29)	(18.94)	(69.53)
Tax assessor (household) variables					
Fair market value ($)	257,589	261,529	259,247	232,816	355,794
	(165,525)	(181,071)	(168,417)	(146,090)	(237,553)
Age of home (years)	20.80	20.75	20.73	15.92	16.84
	(13.17)	(13.45)	(13.37)	(10.98)	(8.63)
Size of property (acres)	0.56	0.57	0.58	0.51	0.62
	(1.05)	(0.97)	(1.09)	(0.78)	(0.92)
Census (neighborhood) variables					
% of people with higher degree	0.72	0.72	0.72	0.68	0.85
	(0.15)	(0.15)	(0.15)	(0.12)	(0.07)
% of people below poverty level	0.04	0.04	0.04	0.04	0.03
	(0.04)	(0.04)	(0.04)	(0.03)	(0.03)
Per capita income	30,559	30,593	30,588	27,263	42,535
	(9079)	(9089)	(9051)	(6984)	(10,617)
% renter-occupied homes	0.09	0.09	0.09	0.09	0.16
	(0.14)	(0.14)	(0.14)	(0.16)	(0.16)
% white	0.83	0.83	0.83	0.81	0.87
	(0.17)	(0.17)	(0.17)	(0.16)	(0.06)

Notes: Means and standard deviations (in parentheses) presented.

Table 2
Nonexperimental replication results using fixed effects panel data estimators.

	(1) Experimental benchmark	(2) Pooled Gwinnett and Fulton comparison	(3) Trimmed sample	(4) Matched without calipers	(5) Matched with calipers
Panel A:					
Social comparison treatment effect	−0.346***	−0.179***	−0.172***	−0.158**	−0.108*
	(0.048)	(0.044)	(0.044)	(0.067)	(0.064)
	[−0.440, −0.252]				
		H_o : (1) = (2)	H_o : (1) = (3)	H_o : (1) = (4)	H_o : (1) = (5)
z-statistic		−2.547	−2.664	−2.272	−2.966
(*p*-value)		(0.011)	(0.008)	(0.023)	(0.003)
Observations	1,347,723	2,362,022	2,347,249	374,977	363,141
Number of households	79,278	140,732	139,863	18,682	18,118
Panel B:					
Technical information treatment effect	−0.012	0.155***	0.165***	0.203***	0.156**
	(0.055)	(0.052)	(0.051)	(0.071)	(0.066)
	[−0.119, 0.096]				
		H_o : (1) = (2)	H_o : (1) = (3)	H_o : (1) = (4)	H_o : (1) = (5)
z-statistic		−2.213	−2.351	−2.387	−1.956
(*p*-value)		(0.027)	(0.019)	(0.017)	(0.050)
Observations	1,346,617	2,360,916	2,346,160	372,807	362,197
Number of households	79,213	140,667	139,799	18,502	17,988

Notes: Robust standard errors clustered at the household level in parentheses. 95% confidence intervals in brackets. Repeated observations in matched samples are taken into account using frequency weights. Caliper width is equal to 1 standard deviation of each covariate. Standard errors are not adjusted for any variation that may be introduced by pre-processing comparison groups.
* $p < 0.1$.
** $p < 0.05$.
*** $p < 0.01$.

with replacement.[6] We apply this matching algorithm with and without calipers; if a treated household does not have a match within the caliper (less than or equal to one standard deviation of each covariate), it is eliminated from the sample. All covariates described in Table 1 serve as matching variables. In our parametric models, repeated matching to the same comparison households is taken into account using frequency weights.

The covariate balance results in Table A.1 (social comparison treatment) and Table A.2 (technical information treatment) corroborate our expectations that trimming and matching improve

[6] We use single nearest-neighbor matching to evaluate Ferraro and Miranda's (2016) framework. Those authors found that one-to-one Mahalanobis matching yielded better covariate balance than propensity score or genetic matching. Other matching algorithms, such as coarsened exact matching, do not offer a straightforward interpretation in a panel context.

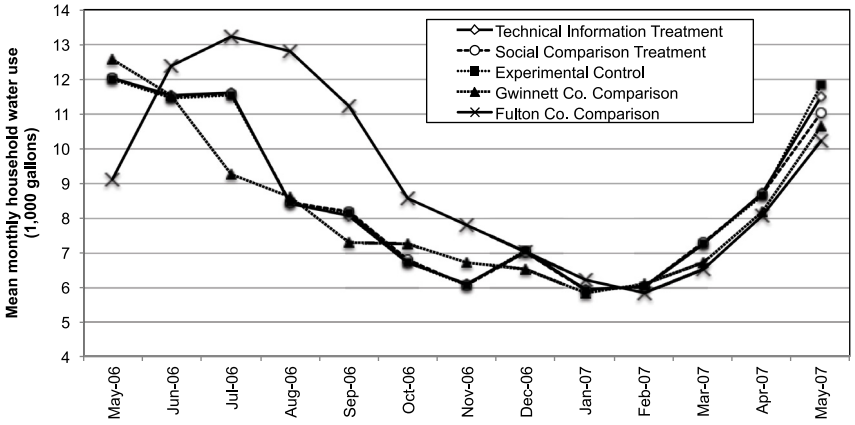
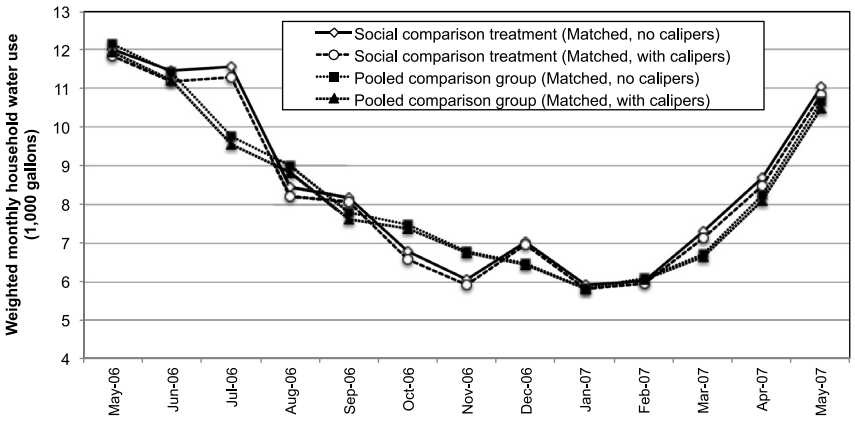
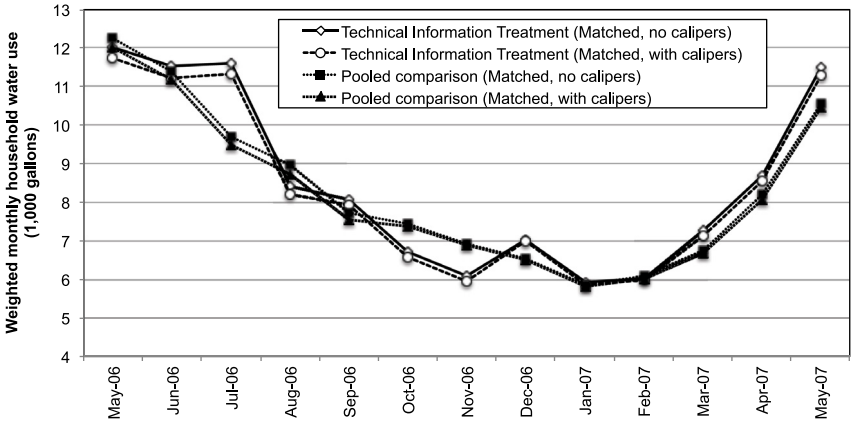


Fig. 1. Pre-treatment mean monthly water consumption (full sample).



(a) Social comparison treatment.



(b) Technical information treatment.

Fig. 2. Pre-treatment weighted mean monthly water (matched samples with and without calipers).

covariate balance and that caliper matching exhibits the best balance.[7]

4.2. Are there parallel trends in pre-treatment consumption?

Prior to presenting empirical results, we consider the assumption of conditional bias stability by evaluating the degree to which pre-treatment trends are similar. In Fig. 1, we plot pre-treatment mean monthly consumption for our treatment and (non-) experimental comparison groups. Although the treatment and comparison trends look identical in the six months prior to treatment, there

[7] For each covariate, we evaluate the improvement in covariate balance in five ways (Lee, 2011): (i) difference in means; (ii) standardized mean difference (Rosenbaum and Rubin, 1985 suggest that a standardized difference greater than 20 should be considered large, although a referee pointed out that other scholars prefer a value of 10); (iii) eQQ mean difference, a non-parametric measure that evaluates rank rather than the precise value of the observations (Ho et al., 2007); (iv) variance ratio between treated and untreated units (Sekhon, 2011); and (v) the Kolmogorov–Smirnov test statistic and bootstrapped p-values using 1000 replications. We do not base our conclusions on p-values alone as they are influenced by sample size, and balance is a quality solely of the sample in question, not as it relates to a population (Ho et al., 2007; Imai et al., 2008).

Table A.1
Covariate balance for social comparison treatment.

		Full sample	Trimmed sample	Matched without calipers	Matched with calipers
Water use May–Oct 2006	Mean difference	−0.092	0.224	1.262	1.117
	Standardized mean difference	−0.226	0.549	3.092	3.073
	Mean raw eQQ difference	1.686	1.636	1.620	1.478
	Variance ratio (Treat/Comp.)	0.884	0.926	1.142	1.071
	Kolmogorov–Smirnov statistic	0.028***	0.027***	0.031***	0.031***
Water use Mar–May 2006	Mean difference	1.314	1.402	0.987	0.871
	Standardized mean difference	11.423	12.17	8.577	8.433
	Mean raw eQQ difference	2.338	1.86	1.082	0.970
	Variance ratio (Treat/Comp.)	0.126	0.467	1.235	1.193
	Kolmogorov–Smirnov statistic	0.094***	0.094***	0.074***	0.075***
Fair market value	Mean difference	−1869.7	−44.0	7646.4	7959.5
	Standardized mean difference	−1.033	−0.024	4.223	5.158
	Mean raw eQQ difference	11746.2	11 184.0	9233.5	9316.7
	Variance ratio (Treat/Comp.)	1.009	1.313	1.210	1.138
	Kolmogorov–Smirnov statistic	0.046***	0.047***	0.033***	0.035***
Age of home (years)	Mean difference	4.601	4.629	0.247	0.185
	Standardized mean difference	34.211	34.487	1.838	1.460
	Mean raw eQQ difference	4.607	4.632	0.571	0.538
	Variance ratio (Treat/Comp.)	1.656	1.641	1.114	1.098
	Kolmogorov–Smirnov statistic	0.171***	0.171***	0.027***	0.027***
Size of property (acres)	Mean difference	0.037	0.03	0.002	−0.002
	Standardized mean difference	3.852	3.982	0.179	−0.433
	Mean raw eQQ difference	0.062	0.054	0.023	0.020
	Variance ratio (Treat/Comp.)	1.435	1.356	1.032	1.054
	Kolmogorov–Smirnov statistic	0.151***	0.152***	0.057***	0.059***
% people with higher education	Mean difference	0.000	0.001	0.001	0.003
	Standardized mean difference	−0.050	0.671	0.642	2.038
	Mean raw eQQ difference	0.019	0.02	0.012	0.011
	Variance ratio (Treat/Comp.)	1.249	1.275	1.164	1.125
	Kolmogorov–Smirnov statistic	0.083***	0.089***	0.066***	0.072***
% people in poverty	Mean difference	0.001	0.001	0.002	0.001
	Standardized mean difference	2.637	3.155	4.205	3.144
	Mean raw eQQ difference	0.005	0.005	0.004	0.003
	Variance ratio (Treat/Comp.)	1.225	1.276	1.172	1.144
	Kolmogorov–Smirnov statistic	0.135***	0.135***	0.091***	0.090***
Per capita income	Mean difference	−501.7	−273.0	581.8	677.1
	Standardized mean difference	−5.519	−3.008	6.401	7.597
	Mean raw eQQ difference	1707.4	1619.5	1004.0	971.8
	Variance ratio (Treat/Comp.)	0.761	0.819	1.091	1.079
	Kolmogorov–Smirnov statistic	0.125***	0.130***	0.102***	0.106***
% renters	Mean difference	−0.017	−0.015	0.008	0.007
	Standardized mean difference	−12.098	−12.089	5.615	5.382
	Mean raw eQQ difference	0.018	0.015	0.008	0.007
	Variance ratio (Treat/Comp.)	0.811	0.761	1.077	1.074
	Kolmogorov–Smirnov statistic	0.070***	0.059***	0.059***	0.058***
% white	Mean difference	0.001	0	−0.004	−0.001
	Standardized mean difference	0.420	0.252	−2.252	−0.333
	Mean raw eQQ difference	0.018	0.018	0.008	0.007
	Variance ratio (Treat/Comp.)	1.351	1.357	1.145	1.104
	Kolmogorov–Smirnov statistic	0.090***	0.088***	0.048***	0.056***

Notes: For each covariate, we evaluate the improvement in covariate balance in five ways (Lee, 2011): (i) difference in means; (ii) standardized mean difference (Rosenbaum and Rubin, 1985 suggest that a standardized difference greater than 20 should be considered large, although a referee pointed out that other scholars prefer a value of 10); (iii) eQQ mean difference, a non-parametric measure that evaluates rank rather than the precise value of the observations (Ho et al., 2007); (iv) variance ratio between treated and untreated units (Sekhon, 2011); and (v) the Kolmogorov–Smirnov equality-of-distributions test statistic and bootstrapped *p*-value using 1000 replications.
*** Indicates significance at the 1% level.

are discrepancies during summer 2006. This difference suggests that Fulton households may not form a good counterfactual for Cobb households. In contrast, Gwinnett households display trends that are similar to the trends of Cobb households before treatment.[8]

In Panels A and B of Fig. 2, we show that, for both treatments, the pre-treatment trend lines become more similar after matching on observed variables in the pooled Fulton and Gwinnett comparison households. Because Gwinnett households are more observationally similar than Fulton households to the treatment groups, Gwinnett observations comprise a larger proportion of the matched sample and are thus weighted more heavily.

5. Experimental and nonexperimental replication results

We assess all nonexperimental estimates according to Ferraro and Miranda's (2016) *Accuracy Criterion*: (a) the nonexperimental

[8] We perform a sensitivity test for matching on pre-treatment water use and present results in Table A.5. Results are qualitatively similar. See Ferraro and Miranda (forthcoming) for a detailed discussion of matching on pre-treatment outcomes with panel data.

Table A.2
Covariate balance for technical information treatment.

		Full sample	Trimmed sample	Matching without calipers	Matching with calipers
	Mean difference	−0.229	0.068	1.460	1.240
	Standardized mean difference	−0.576	0.171	3.670	3.484
Water use May–Oct 2006	Mean raw eQQ difference	1.796	1.811	1.803	1.604
	Variance ratio (Treat/Comp.)	0.840	0.860	1.105	1.070
	Kolmogorov–Smirnov statistic	0.028***	0.028***	0.030***	0.031***
	Mean difference	1.313	1.417	1.030	0.917
	Standardized mean difference	11.288	12.169	8.856	8.505
Water use Mar–May 2006	Mean raw eQQ difference	2.365	1.901	1.158	1.056
	Variance ratio (Treat/Comp.)	0.129	0.476	1.295	1.259
	Kolmogorov–Smirnov statistic	0.095***	0.095***	0.077***	0.076***
	Mean difference	−5809.9	−3474.0	7526.9	8169.7
	Standardized mean difference	−3.510	−2.093	4.547	5.575
Fair market value	Mean raw eQQ difference	10 260.4	8480.6	8571.0	8910.1
	Variance ratio (Treat/Comp.)	0.843	1.104	1.116	1.136
	Kolmogorov–Smirnov statistic	0.041***	0.042***	0.037***	0.039***
	Mean difference	4.645	4.657	0.196	0.125
	Standardized mean difference	35.270	35.444	1.487	0.999
Age of home (years)	Mean raw eQQ difference	4.657	4.666	0.544	0.516
	Variance ratio (Treat/Comp.)	1.588	1.571	1.111	1.095
	Kolmogorov–Smirnov statistic	0.176***	0.175***	0.029***	0.028***
	Mean difference	0.027	0.012	0.000	−0.002
	Standardized mean difference	2.548	1.961	0.015	−0.502
Size of property (acres)	Mean raw eQQ difference	0.059	0.047	0.023	0.018
	Variance ratio (Treat/Comp.)	1.674	0.915	0.957	1.068
	Kolmogorov–Smirnov statistic	0.156***	0.157***	0.058***	0.058***
	Mean difference	0.000	0.001	0.001	0.003
	Standardized mean difference	0.141	0.822	0.718	2.086
% people with higher education	Mean raw eQQ difference	0.019	0.020	0.012	0.011
	Variance ratio (Treat/Comp.)	1.247	1.274	1.152	1.103
	Kolmogorov–Smirnov statistic	0.085***	0.093***	0.070***	0.074***
	Mean difference	0.001	0.002	0.001	0.001
	Standardized mean difference	3.694	4.347	3.975	2.963
% people in poverty	Mean raw eQQ difference	0.006	0.006	0.004	0.003
	Variance ratio (Treat/Comp.)	1.241	1.282	1.150	1.128
	Kolmogorov–Smirnov statistic	0.132***	0.131***	0.088***	0.088***
	Mean difference	−535.465	−263.000	551.548	638.022
	Standardized mean difference	−5.898	−2.897	6.075	7.190
Per capita income	Mean raw eQQ difference	1721.7	1612.2	983.2	930.2
	Variance ratio (Treat/Comp.)	0.760	0.833	1.096	1.070
	Kolmogorov–Smirnov statistic	0.124***	0.130***	0.096***	0.100***
	Mean difference	−0.017	−0.013	0.007	0.006
	Standardized mean difference	−12.605	−10.413	5.305	4.893
% renters	Mean raw eQQ difference	0.018	0.013	0.008	0.007
	Variance ratio (Treat/Comp.)	0.768	0.790	1.067	1.068
	Kolmogorov–Smirnov statistic	0.063***	0.051***	0.058***	0.055***
	Mean difference	0.002	0.001	−0.003	−0.001
	Standardized mean difference	1.425	0.890	−1.981	−0.488
% white	Mean raw eQQ difference	0.017	0.017	0.008	0.007
	Variance ratio (Treat/Comp.)	1.326	1.337	1.134	1.091
	Kolmogorov–Smirnov statistic	0.086***	0.083***	0.054***	0.061***

Notes: For each covariate, we evaluate the improvement in covariate balance in five ways (Lee, 2011): (i) difference in means; (ii) standardized mean difference (Rosenbaum and Rubin, 1985 suggest that a standardized difference greater than 20 should be considered large, although a referee pointed out that other scholars prefer a value of 10); (iii) eQQ mean difference, a non-parametric measure that evaluates rank rather than the precise value of the observations (Ho et al., 2007); (iv) variance ratio between treated and untreated units (Sekhon, 2011); and (v) the Kolmogorov–Smirnov equality-of-distributions test statistic and bootstrapped *p*-value using 1000 replications.
*** Indicates significance at the 1% level.

point estimate should be in the 95% confidence interval of the experimental point estimate; (b) the correct inference should be made when testing the null hypothesis of no treatment effect (type 1 error = 5%). Additionally, we include a cross-equation test of statistical significance between the experimental benchmark and the nonexperimental estimate for each model.[9]

Following the guidance from Ho et al. (2007), we base all of our statistical inference on estimated variances without adjusting for any variation introduced by the pre-processing procedure.

Because Cobb County households appear to be more similar to Gwinnett residents than Fulton residents (Tables 1, A.1, and A.2), we first present results that treat each of the comparison counties as distinct populations. The details of the results are presented, for brevity, in Tables A.3 and A.4. Using Fulton as the only source for comparison units and pre-processing the data with caliper matching, the non-experimental social comparison treatment effect estimate meets the *Accuracy Criterion*. The nonexperimental technical information treatment effect estimate just misses

[9] We do not rely solely on statistical significance across estimated parameters, which depends on the precision of the nonexperimental estimator and the sample size, because we do not want to infer that the nonexperimental design performs well simply because the estimate has a large confidence interval.

Table A.3
Nonexperimental estimates using Fulton County as only comparison group.

	(1) Experimental benchmark	(2) Pooled comparison	(3) Trimmed Sample	(4) Matched without calipers	(5) Matched with Calipers
Panel A:					
Social comparison treatment effect	−0.346***	−1.007***	−0.979***	−0.493***	−0.416***
	(0.048)	(0.083)	(0.112)	(0.150)	(0.134)
	[−0.440, −0.252]				
		H_o : (1) = (2)	H_o : (1) = (3)	H_o : (1) = (4)	H_o : (1) = (5)
z-statistic		6.88	5.179	0.929	0.492
(*p*-value)		(<0.001)	(<0.001)	(0.353)	(0.623)
Observations	1,347,723	745,909	496,468	378,165	248,693
Number of households	79,278	43,877	29,204	14,477	9945
Panel B:					
Technical information treatment effect	−0.012	−0.673***	−0.603***	−0.204	−0.135
	(0.055)	(0.087)	(0.112)	(0.157)	(0.133)
	[−0.119, 0.096]				
		H_o : (1) = (2)	H_o : (1) = (3)	H_o : (1) = (4)	H_o : (1) = (5)
z-statistic		6.415	4.736	1.16	0.861
(*p*-value)		(<0.001)	(<0.001)	(0.246)	(0.389)
Observations	1,346,617	744,803	506,820	376,039	248,692
Number of households	79,213	43,812	29,813	14,420	9972

Notes: Robust standard errors clustered at the household level in parentheses. 95% confidence intervals in brackets. Repeated observations in matched samples are taken into account using frequency weights. Caliper width is equal to 1 standard deviation of each covariate. Standard errors are not adjusted for any variation that may be introduced by pre-processing comparison groups. Our estimates diverge slightly from the estimates in Ferraro and Miranda (forthcoming) due to independent merging of household and socioeconomic data with water billing records, and the corresponding difference in matches. The estimates presented in this paper exhibit a larger match success rate that Ferraro and Miranda.
*** $p < 0.01$.

satisfying the criterion, but in the cross-equation test we cannot reject the null hypothesis that the estimated effect is equal to the experimental benchmark. These results are consistent with Ferraro and Miranda's conclusions. In contrast, using Gwinnett County as the only source for comparison households, we fail to satisfy any accuracy criteria for both treatments, despite having improved covariate balance and common pre-treatment trends.

Table 2 presents our comparisons of the experimental and non-experimental estimates, pooling Gwinnett and Fulton households. In Panel A, we show the estimates for the social comparison treatment effect. In column (1), the experimental benchmark for this treatment is −0.346. In other words, households treated with a social comparison message reduced consumption by 346 gallons per month, on average. Column (2) presents the nonexperimental estimate using the full sample of Fulton and Gwinnett households as comparison groups. Like the experimental estimate, the nonexperimental estimate is negative and statistically significant. Yet it is less than one-third the magnitude of the experimental estimate.

The trimmed sample in column (3) does not perform any better despite being more balanced on observables. The matched samples without and with calipers, in columns (4) and (5), do not perform better than the pooled sample. The matched sample with calipers was the most balanced across treatment and comparison groups and implies a nonexperimental treatment effect of −0.108. While the non-experimental estimates have the same sign as the experimental benchmark estimate, they fall outside its 95% confidence interval (*Accuracy Criterion*) and are statistically different from the benchmark (*p*-value = 0.01). This result stands in contrast to Ferraro and Miranda (forthcoming) – as well as our replication in Table A.3 – who replicate the experimental benchmark with caliper matching and only Fulton households in the comparison group.

Results for the technical information treatment are presented in Panel B of Table 2. The experimental benchmark in the first column is −0.012, a small and statistically insignificant response to treatment. The caliper-matched sample provides the best balance, as shown in Table A.2, as well as the smallest treatment effect out of the pre-processed samples. The estimated effect, however, is positive and statistically significant, it falls outside the 95% confidence interval of the experimental benchmark (*Accuracy Criterion*), and it is statistically different from the experimental estimate (*p*-value = 0.05). This result also contradicts the main findings from Ferraro and Miranda (forthcoming), who replicate the experimental benchmark with caliper matching (excluding Gwinnett households).[10] Results for both treatments are robust to excluding pre-treatment water use from the set of matching covariates, as well as limiting the caliper width to 0.5 and 0.25 standard deviations (see Table A.5).

6. Concluding remarks

These results remind us that in observational settings the choice of an appropriate comparison group in the spirit of Heckman et al. (1997) is challenging. Ferraro and Miranda (2014) contend that pre-processing data to make treatment and comparison groups observationally more similar in pre-treatment characteristics and trends results in an observational design more likely to replicate an experimental benchmark. However, after we enlarge the pool of potential comparison units, which improves both covariate balance and parallel pre-treatment trends, the FEPD estimator performs worse. This result reminds researchers that indirect tests of untestable identification assumptions (e.g., parallel pre-treatment trends) are no guarantee that the assumptions are satisfied. Further, this study highlights the dangers of depending on single studies for evidence about program impacts and the importance of replication in the social sciences and program evaluation.

[10] We include sensitivity tests in Table A.3 with calipers equal to 0.5 SD and 0.25 SD of each covariate. Results for the social comparison treatment do not improve. We can, however, replicate the experimental benchmark according to our *Accuracy Criterion* for the information treatment (panel B) with smaller caliper widths.

Table A.4
Nonexperimental estimates using Gwinnett County as only comparison group.

	(1) Experimental benchmark	(2) Pooled comparison	(3) Trimmed sample	(4) Matched without calipers	(5) Matched with calipers
Panel A:					
Social comparison treatment effect	−0.346***	0.105**	0.091**	−0.019	−0.001
	(0.048)	(0.041)	(0.041)	(0.067)	(0.064)
	[−0.440, −0.252]				
		H_o : (1) = (2)	H_o : (1) = (3)	H_o : (1) = (4)	H_o : (1) = (5)
z-statistic		−7.105	−6.900	−3.945	−4.309
(*p*-value)		(<0.001)	(<0.001)	(<0.001)	(<0.001)
Observations	1,347,723	1,805,187	1,764,950	374,439	353,899
Number of households	79,278	107,977	105,565	18,831	17,830
Panel B:					
Technical information treatment effect	−0.012	0.439***	0.427***	0.361***	0.308***
	(0.055)	(0.049)	(0.049)	(0.073)	(0.068)
	[−0.119, 0.096]				
		H_o : (1) = (2)	H_o : (1) = (3)	H_o : (1) = (4)	H_o : (1) = (5)
z-statistic		−6.125	−5.955	−4.068	−3.67
(*p*-value)		(<0.001)	(<0.001)	(<0.001)	(<0.001)
Observations	1,346,617	1,804,081	1,757,422	372,282	352,944
Number of households	79,213	107,912	105,117	18,649	17,720

Notes: Robust standard errors clustered at the household level in parentheses. 95% confidence intervals in brackets. Repeated observations in matched samples are taken into account using frequency weights. Caliper width is equal to 1 standard deviation of each covariate. Standard errors are not adjusted for any variation that may be introduced by pre-processing comparison groups.
** $p < 0.05$.
*** $p < 0.01$.

Table A.5
Sensitivity for nonexperimental replication results using fixed effects panel data estimators.

	(1) Experimental benchmark	(2) Matched with calipers *No pre-treatment water use*	(3) Matched with calipers *Caliper width*: 0.5 *SD*	(4) Matched with calipers *Caliper width*: 0.25 *SD*
Panel A:				
Social comparison treatment effect	−0.346***	−0.150	−0.049	−0.101
	(0.048)	(0.096)	(0.066)	(0.097)
	[−0.440, −0.252]			
		H_o : (1) = (2)	H_o : (1) = (3)	H_o : (1) = (4)
z-statistic		−1.82	−3.634	−2.257
(*p*-value)		(0.069)	(<0.001)	(0.024)
Observations	1,347,723	368,887	276,986	70,156
Number of households	79,278	16,726	13,966	3,670
Panel B:				
Technical information treatment effect	−0.012	0.219**	0.159**	−0.185*
	(0.055)	(0.110)	(0.069)	(0.109)
	[−0.119, 0.096]			
		H_o : (1) = (2)	H_o : (1) = (3)	H_o : (1) = (4)
z-statistic		−1.188	−1.933	−1.422
(*p*-value)		(0.061)	(0.053)	(0.155)
Observations	1,346,617	368,314	274,515	66,597
Number of households	79,213	16,530	13,827	3478

Notes: Robust standard errors clustered at the household level in parentheses. 95% confidence intervals in brackets. Repeated observations in matched samples are taken into account using frequency weights. Caliper width is equal to 1 standard deviation of each covariate in column (2), 0.5 in column (3), and 0.25 in column (4). Standard errors are not adjusted for any variation that may be introduced by pre-processing comparison groups.
* $p < 0.1$.
** $p < 0.05$.
*** $p < 0.01$.

Acknowledgments

For feedback and comments, the authors thank Philip Gleason and participants at the 2014 Association for Public Policy Analysis and Management conference. For the water data, the authors thank Kathy Nguyen, Herb Richardson, and Kathleen Brown of Cobb County Water System, Diane Raymond of Fulton County, and Alisha Voutas, Gwinnett County Department of Water Resources.

Appendix A

See Tables A.1–A.5 for additional results.

Appendix B. Supplementary data

Supplementary material related to this article can be found online at http://dx.doi.org/10.1016/j.econlet.2016.11.029.

References

Angrist, Joshua, Krueger, Alan, 1999. Empirical strategies in labor economics. In: Ashenfelter, Orley, Card, David (Eds.), Handbook of Labor Economics, Vol. 3. pp. 1277–1366. (Chapter 23).

Cook, T., Shadish, W., Wong, V., 2008. Three conditions under which observational studies produce the same results as experiments. J. Policy Anal. Manag. 274, 724–750.

Crump, R., Hotz, J., Imbens, G., Mitnik, O., 2009. Dealing with limited overlap in estimation of average treatment effects. Biometrika 96 (1), 187–199.

Dehejia, Rajeev, 2005. Practical propensity score matching: A reply to smith and todd. J. Econometrics 125, 355–364.

Ferraro, P.J., Miranda, J.J., 2014. The performance of non-experimental designs in the evaluation of environmental programs: A design-replication study using a large-scale randomized experiment as a benchmark. J. Econ. Behav. Organ. 107, 344–365.

Ferraro, P.J., Miranda, J.J., 2016. Panel data designs and estimators as substitutes for randomized controlled trials in the evaluation of public programs. J. Assoc. Environ. Resour. Econom. (forthcoming).

Ferraro, P.J., Price, M.-K., 2013. Using non-pecuniary strategies to influence behavior: Evidence from a large-scale field experiment. Rev. Econ. Stat. 95 (1), 64–73.

Heckman, J.J., Smith, J.A., 1995. Assessing the case for social experiments. J. Econ. Perspect. 9 (2), 85–110.

Heckman, J.-J., Ichimura, H., Todd, P., 1997. Matching as an econometric evaluation estimator: Evidence from evaluating a job training program. Rev. Econom. Stud. 64 (4), 605–654.

Ho, D., Imai, K., King, G., Stuart, E., 2007. Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. Polit. Anal. 15, 199–236.

Imai, Kosuke, King, Gary, Stuart, Elizabeth A., 2008. Misunderstandings between experimentalists and observationalists about causal inference. J. Roy. Statist. Soc. Ser. A 171 (2), 481–502.

Lalonde, R., 1986. Evaluating the econometric evaluations of training with experimental data. Amer. Econ. Rev. 76, 604–620.

Lee, W.-S., 2011. Propensity score matching and variations on the balancing test. Empir. Econom. 1–34. 26 May 2011.

Rosenbaum, P., Rubin, D., 1985. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. Amer. Statist. 39 (1), 33–38.

Sekhon, J., 2011. Multivariate and propensity score matching software with automated balance optimization: The matching package for R. J. Stat. Softw. 42 (7), 1–52.

Smith, J., Todd, P., 2005. Does matching overcome Lalonde's critique of nonexperimental estimators? J. Econometrics 125, 305–353.

Wichman, C.J., Taylor, L.O., von Haefen, R.H., 2016. Conservation policies: Who responds to price and who responds to prescription? J. Environ. Econ. Manag. 79, 114–134.