European Economic Review 144 (2022) 104094

Contents lists available at ScienceDirect

European Economic Review

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

Information campaigns for residential energy conservation

Mark A. Andor [a,*], Andreas Gerster [b], Jörg Peters [a,c]

[a] RWI – Leibniz Institute for Economic Research, Germany
[b] University of Mannheim, L7 3-5, 68161 Mannheim, Germany
[c] University of Passau, Germany

ARTICLE INFO

JEL classification:
D12, D83, L94, Q41

Keywords:
Imperfect information
Information letters
Behavioral public economics
Energy efficiency
Energy conservation
Non-price interventions
Targeting

ABSTRACT

This paper evaluates an intervention that randomized information letters about energy efficient investments and behaviors among 120,000 customers of two utilities in Germany. We find that conservation effects differ considerably between both utilities, ranging from a precisely estimated zero effect to −1.4%. By contrast, we do not detect significant framing effects from presenting savings in monetary or ecological terms. Based on random causal forest methods, we show that the effect heterogeneity across utilities cannot be explained by socio-demographic characteristics. Our results demonstrate the importance of site-specific factors for the effectiveness of information campaigns, which has crucial implications for targeting and the ability to infer population-wide effect sizes from pilot studies.

1. Introduction

Individuals are often not fully informed when making decisions (Stigler, 1961). Information provision has been shown to affect individual decision-making in various contexts, including social-benefit take-up, agriculture, health, and water conservation (Bhargava and Manoli, 2015; Duflo and Saez, 2003; Ferraro and Price, 2013; Hanna et al., 2014; Jalan and Somanathan, 2008). A widespread policy tool used against imperfect information are campaigns that aim to improve households' decision-making by closing knowledge gaps. To reach ambitious energy conservation goals, for instance, many governments have implemented programs that inform consumers about effective energy-saving behaviors and investments, such as the *Energy Efficiency Awareness Program* in Canada, *Action for Warm Homes* in the United Kingdom, and *Energy Action at Home* in the United States. Information campaigns are appealing from a practical perspective as they do not rely on expensive technology. Yet, despite their widespread use, evidence on the effectiveness of large-scale campaigns is scarce and inconclusive.[1]

This paper tests the impact of a letter-based information campaign about the savings potential of energy efficient behaviors and investments in a population of more than 120,000 households in Germany. In our randomized controlled trial (RCT), conducted with two utilities, households in our treatment groups received four information letters within one year.[2] We test the effectiveness of different framings that are commonly employed in information campaigns on sustainable behavior. Specifically, we implement three treatment arms in which savings information is displayed in (i) monetary terms, (ii) in terms of carbon dioxide (CO_2) emission

* Correspondence to: RWI – Leibniz Institute for Economic Research, Hohenzollernstr. 1-3, D-45128 Essen, Germany.
E-mail address: andor@rwi-essen.de (M.A. Andor).

[1] In the context of energy conservation, studies on the impact of energy-saving information typically rely on small samples and find largely different effect sizes (−12 to 8% in Delmas et al., 2013 and −17 to 5% in Buckley, 2020).

[2] This experiment was not pre-registered. When we designed the study in 2014, pre-specification of RCTs was not as widespread as it is today (see Ofosu and Posner, 2019, 2020). We furthermore saw little necessity to tie our hands because we anticipated very limited leeway in the analysis of our data, as we also outline in Section 2.3. Whenever we use rich secondary data, we explicitly declare this analysis to be explorative. Yet, we do acknowledge the advantage of clearly formulated pre-analysis plans, even in straightforward experiments with limited data availability.

https://doi.org/10.1016/j.euroecorev.2022.104094
Received 19 March 2021; Received in revised form 5 January 2022; Accepted 22 January 2022
Available online 12 March 2022

reductions, and (iii) a combination of both. We estimate the treatment effect of our intervention on annual electricity consumption and investigate its persistence over two consecutive years. Furthermore, we explore treatment effect heterogeneity and test how machine learning algorithms for treatment effect prediction can be exploited for improving the cost-effectiveness of informational interventions through optimal targeting.

We refer to our intervention as an *information campaign* because it aims to improve households' knowledge about the consequences of their energy-related behaviors and investments.[3] Our letters contain energy saving tips that we selected in cooperation with the *Verbraucherzentrale NRW*, Germany's largest nonprofit organization for consumer protection, and the *Energieagentur.NRW*, a government-funded agency to promote energy efficiency. To ensure that consumers understand and appreciate the campaign, we designed the letters with a marketing consultancy. We also hired a market research institute to conduct qualitative pretests on our letters before we started the field test. We varied the framing of the letters because the importance of framing effects has been emphasized by numerous contributions in the field of economic psychology (e.g., Kahneman and Tversky, 1979; Tversky and Kahneman, 1981; Levin et al., 1998). In our context, previous studies suggest that framing the savings potential in monetary versus environmental terms could influence the effectiveness of energy conservation campaigns (Bolderdijk et al., 2013; Asensio and Delmas, 2015).

We implemented our study in cooperation with two utilities; one large supra-regional utility (henceforth, SREG), with a customer base covering wide parts of both rural and urban Germany, and a smaller regional utility (henceforth, REG), that operates in the mostly rural north-eastern part of Germany. This setting allows us to test the effectiveness of an identical information campaign in two sample sites. Our study can be classified as a natural field experiment (Levitt and List, 2009), as households were not informed about the study and no survey was implemented. We retrieved household level information including electricity consumption from the utilities' customer database and annual metering. In addition, we obtained data on household electricity consumption for a full year after the intervention had ended, which allows us to estimate the persistence of treatment effects. We did not pre-register the experiment, but the analyses underlying our main results closely follow our experimental design, as discussed in Section 2.3. Once we probe into subgroups and use secondary data on households' socio-demographic characteristics, we transparently label these additional heterogeneity analyses as exploratory.

We find that the average treatment effect in the year of the treatment ranges from a precisely estimated zero effect for SREG customers (−0.06%) to −1.36% for REG customers, and attenuates by about 27% in the year after the treatment had ended. The effectiveness of the intervention is thus limited when applied to the entire population of residential electricity users, but differs significantly across utilities. By contrast, we do not detect statistically significant differences in the effect sizes of the different framings. Furthermore, we explore treatment effect heterogeneity across utilities based on comprehensive socio-demographic data. We find that customers with large baseline usage have larger treatment effects at REG, but not at SREG. To test whether observable covariates can explain the difference in treatment effects, we employ a random causal forest machine learning algorithm developed by Wager and Athey (2018). These methods have been developed to capture even complex treatment effect heterogeneity patterns without being susceptible to data mining issues (e.g., Athey and Imbens, 2016). Our results show that the differences in observable characteristics cannot explain the differences in treatment effects across utilities.

Our study contributes to the literature in four ways. First, it relates to studies that evaluate the effectiveness of information campaigns as a policy instrument. Despite the widespread use of such interventions, evidence on their effectiveness is, so far, inconclusive. For example, previous studies have shown that employees who have received letters about their expected pensions save more for retirement in Germany (Dolls et al., 2018), but not in the U.S. (Carter and Skimmyhorn, 2018). In the context of water use, the effect of providing conservation tips varies widely across studies, from 1% (Ferraro and Price, 2013) up to 2 to 5% (Goette et al., 2019; Tonke, 2020). Even more drastically, studies on information provision to conserve energy have found effect sizes ranging from less than −10 to about 8% (see e.g. Buckley, 2020; Delmas et al., 2013 and Appendix Table A.1 for an overview).[4] We contribute to this literature by implementing an RCT with more than 120,000 participants. The large sample size and hence our high statistical power enable us to detect even small conservation effects. Underpowered and non-experimental study designs have been identified as potential reasons for the highly heterogeneous findings from energy conservation interventions (e.g., Andor and Fels, 2018; Delmas et al., 2013; Karlin et al., 2015). In addition, we evaluate our intervention in two sample sites and can thus test for site-specific effects. We find that treatment effects vary substantially across sites, which may explain some of the effect heterogeneity found in earlier studies.

Second, we contribute to a growing body of literature on the scalability and external validity of interventions (e.g., Al-Ubaydli et al., 2017a,b; Allcott, 2015; Dehejia et al., 2019; Vivalt, 2020). Previous research has shown that effect sizes often diminish when interventions are brought to scale for a variety of reasons, such as site and partner selection, differences in program implementation, and in the composition of pilot and target populations (Al-Ubaydli et al., 2017a,b; Allcott, 2015). In our study, we find that effect sizes differ considerably across utilities although both utilities took part in our study at the same point in time, the information letters had exactly the same content, and program implementation was almost indistinguishable for both utilities. Our finding that the effectiveness of information campaigns varies substantially across sample sites demonstrates the difficulty to generalize findings in that context (see, e.g., Vivalt, 2020 for a discussion of the generalizability of impact evaluations more broadly). Furthermore, we use detailed socio-demographic data and a causal forest machine learning methodology developed by Wager and Athey (2018)

[3] This distinguishes our intervention from behavioral interventions, such as social comparisons (e.g., Allcott, 2011; Jaime Torres and Carlsson, 2018) and feedback (e.g., Tiefenbeck et al., 2013, 2018).

[4] In addition, evidence from RCTs shows that seller-provided information about the fuel economy of cars and the energy efficiency of appliances does not affect purchase decisions (Allcott and Knittel, 2017; Allcott and Sweeney, 2017).

to show that observable characteristics cannot explain effect heterogeneity across utilities. In our study, evidence from one site is virtually uninformative about the effectiveness at the other site. This evidence contrasts with earlier studies that have found that effect sizes in one site at least partially predict the effect sizes in other sites (e.g., Meager, 2019; Allcott, 2015) and supports the importance of "macro covariates" (Dehejia et al., 2019). Our finding implies that it is particularly difficult to draw conclusions about the population-wide effects based on evidence from few, or even one, pilot studies in the context of information campaigns.

Third, we analyze the potential of targeting to increase the cost-effectiveness of large-scale interventions. In principle, targeting may be particularly important for information provision because some individuals engage less in a beneficial behavior after learning about lower-than-expected benefits (Byrne et al., 2018; Schultz et al., 2007; Wichman, 2017). This rationale accords with studies in the context of social-comparison based reports that have found substantial welfare benefits from targeting (Allcott and Kessler, 2019; Knittel and Stolper, 2019). In contrast to these studies, we find that targeting plays only a limited role in the context of information campaigns. For REG, we detect that targeting households with a large baseline consumption would increase the cost-effectiveness of the intervention. However, this pattern does not extend to SREG, where we do not find any sizable effect heterogeneity. The fact that heterogeneity patterns differ across sites requires site-specific evidence from pilot studies in order to define targeting strategies, which may be difficult to obtain in practice.

Fourth, we contribute to the literature that has evaluated the effectiveness of behavioral interventions for resource conservation. Many interventions, such as social-comparison based home energy reports (HER), contain a multitude of elements, including a social comparison module, consumption feedback, and electricity-saving tips. We isolate the effectiveness of electricity-saving tips by evaluating information letters that only contain that element. In our study, the conservation effect reaches 1.4% for one utility. Hence, electricity-saving tips might partly explain the effect sizes of 0.5–3.3% (e.g., Allcott, 2011, 2015) that have been found for HER interventions.

The paper is structured as follows. In the next section, we describe the experimental design and the data. In Section 3, we investigate the average conservation effect of the information campaign and explore treatment effect heterogeneity. Section 4 discusses the implications of our results for scalability and the optimal targeting of informational interventions. Section 5 concludes.

2. Treatment design, implementation and data

2.1. Design of information letters (IL)

A pivotal element of our study is the content of the information letters. In an intense preparatory phase, we cooperated with the two energy utilities that implemented the intervention, as well as with two energy efficiency advocacy agencies: *Verbraucherzentrale NRW*, Germany's largest non-profit organization for consumer protection, and *Energieagentur.NRW*, a government-funded agency to promote energy efficiency. Furthermore, we hired the marketing consultancy *brandseven* to design the letters and the market research institute *Rheingold – Institute for Qualitative Market and Media Research* to conduct qualitative pre-tests of our letters.

In a first step, and in cooperation with experts from *Verbraucherzentrale NRW* and *Energieagentur.NRW*, we collected all possible tips for energy efficient investments and behaviors that may apply to typical German households from consumer protection agencies, product testing companies, and governmental agencies (for our sources, see, e.g., Appendix Table A2). We conducted a qualitative assessment of the tips based on five criteria that we evaluated using a three-point Likert scale. The criteria included the size of the potential energy savings that can be realized (*impact*), the share of the population that the tip applies to (*relevance*), the level of technical understanding required to understand the tip (*intelligibility*), and the financial implementation cost (*financial cost*), as well as the non-financial *implementation effort* (for details, see Appendix Section A3).

Afterwards, we selected the energy saving tips with the highest average score and designed appealing and easily understandable letters in cooperation with *brandseven*, taking into account their experience with the customers of electricity providers. In a second step, we partnered with *Rheingold* to pre-test the letters and the selection of energy saving tips. To this end, *Rheingold* recruited 16 volunteers for qualitative in-depth interviews using a non-random sampling approach in downtown Munich. In spite of the small sample size, the participants represented all relevant customer groups in terms of sex, age, housing conditions (apartment vs. single-family homes, owning property vs. renting), marital status and family size. *Rheingold* encouraged participants to describe in detail what they perceived and thought when reading the letters. Based on their responses, *Rheingold* created a typology of three groups of energy savers: first, those who are eager to save energy out of conviction to reduce their environmental footprint; second, those who are eager to reduce their electricity bill for monetary reasons; and third, those who believe in technology and consider energy saving more as an opportunity to invest in modern appliances. *Rheingold* documented what they refer to as "turning points and psychological transitions" of interviewees and came up with suggestions for the most impactful saving tips.

Using these insights, we fine-tuned our selection of savings tips and designed three treatment groups to test for framing effects. We translated the kWh savings into monetary terms for the first treatment group (*monetary framing* in the following), into CO_2 savings for the second group (*ecological framing*), and we combined both dimensions for the third group (*combined framing*). The letters in all treatment groups were identical except for the differences in the framing. As a final consolidation step, we again shared and discussed the drafts of the letters with experts from *Verbraucherzentrale NRW* and *Energieagentur.NRW*. In Appendix Section A2 and A3, we show screenshots of the letters, describe all selected tips, and detail our calculations of their savings potential.

The first letter focused on hot water usage and the bathroom, the second on cooking and kitchen appliances, the third on lighting and the living room, and the fourth on entertainment and communication devices. Each of the four letters presented two investment tips and two behavioral tips. Both utilities sent letters of identical content and only adjusted the presentation to match their corporate design. Figs. 1(a) and 1(b) present examples for one of our energy saving tips in the ecological and monetary framing,

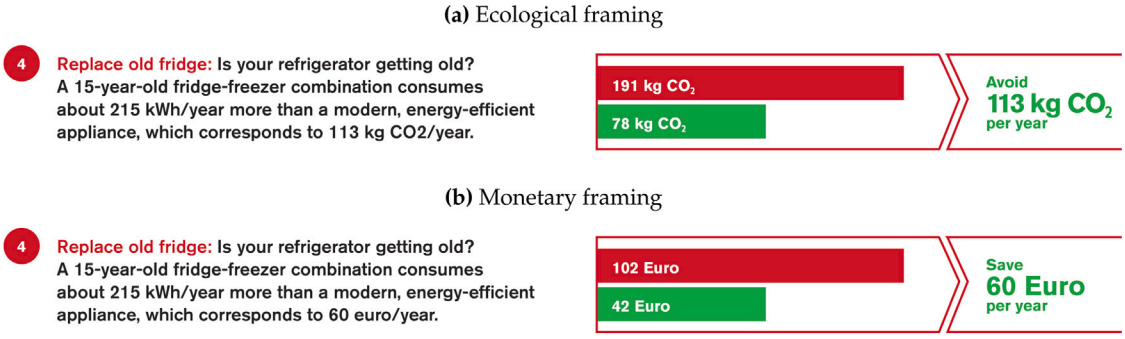

Fig. 1. Example for the presentation of electricity-saving tips (English translation). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

respectively. Every tip is associated with a brief explanation and the yearly kWh savings. The displayed tip proposes to replace an old refrigerator with a new energy-efficient one, which results in a reduction in annual operating cost of around 60 EUR and a reduction in CO_2 emissions of 113 kg. The bars visualize the annual cost (monetary framing) or CO_2 emissions (ecological framing) prior to implementing the tip (red bar) and thereafter (green bar), as well as the total savings. In addition to one page that contained the energy-saving tips, our mailing also included a cover letter that introduced the household to the campaign and a reminder card that summarized the saving tips (for details, see Appendix A2).

2.2. Implementation of the randomized controlled trial and data

For the implementation of the randomized intervention, we cooperated with one regional and one supra-regional utility, both located in Germany. The regional utility (henceforth REG) has around 160,000 customers mainly in the North-East of Germany, whereas the supra-regional utility (henceforth SREG) provides electricity to more than six million customers all over Germany, with a focus on the West and South-East of Germany. Out of this population, we use a sample of around 123,000 residential electricity consumers in total; 115,000 from SREG and 8000 from REG. We randomized the intervention among those households that received their annual electricity bill between mid-August and the end of October 2014. The randomization was stratified by the households' baseline electricity use and the utility. Due to the smaller sample size, we did not implement the *combined framing* treatment in the REG sample so that we tested all three treatments in the SREG sample and the *ecological framing* and *monetary framing* in the REG sample.

The four letters were sent to households on a quarterly basis. We sent the first information letter shortly after a household had had its yearly meter reading and had also received the electricity bill for 2014, which constitutes the baseline year of our analysis. Hence, at the time of the next annual metering in 2015, households would have been exposed to the full four letter treatment for about three months. After receiving their 2015 electricity bill, households did not receive any further letters. We observe the electricity consumption for another year, until households received their 2016 electricity bill. This additional year allows us to analyze how treatment effects evolve over time. One might, for example, expect that tips to invest into more energy efficient appliances are realized only after some time. By contrast, behavioral responses might be stronger immediately after the reception of the letter and then attenuate over time.

As households were not interviewed or informed about participation in an experiment, we can rule out biases through survey, John Henry, and Hawthorne effects (see, e.g., Schwartz et al., 2013). Our sample includes only households that had been with the electricity supplier for at least one year, in order to draw on baseline consumption data. Beyond electricity consumption, the only information we received from the two utilities is the consumer's tariff and address. For data protection reasons, the address information we obtained included the zip code and street name, but not the house number. Since the liberalization of the electricity market in 1998, customers in Germany have been able to choose freely between the various electricity providers and tariffs. However, many households have never switched their tariff. We refer to this group as the "default" tariff group. Moreover, providers usually offer additional tariffs that differ with respect to price and non-price features. SREG, for example, offers a so-called "green" tariff, for which it promises to feed-in an amount of electricity from renewable energy sources equivalent to the customer's consumption. At REG, all customers receive electricity from renewable sources, but customers can opt for a tariff that promises additional investments in climate change projects. We code these REG customers as "green" because they make an active pro-environmental choice among the tariffs offered by that utility. Furthermore, both utilities offer a heating electricity tariff, where electricity is separately metered for peak- and off-peak times of day, which then allows households to operate electric storage heaters that absorb heat overnight and release it during the day.

Based on the households' addresses, we merged our data set with information at the 1 km grid-level that we obtained from a socio-demographic data provider, *microm* (microm, 2015). This data set includes population densities, unemployment rates, the average purchasing power per household, the percentage of retirees, and the percentage of foreign household heads. Our data correspond to averages at a 1 km grid-level and thus measures the households' socio-demographic status with error. In principle,

Table 1
Balance of baseline characteristics between experimental groups.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	SREG					REG			
	Control	Monetary	Ecol.	Combined	P-value	Control	Monetary	Ecol.	P-value
Baseline cons., in kWh per day	9.05	9.05	9.01	9.07	(0.88)	7.72	7.90	7.86	(0.34)
Default tariff, in %	0.90	0.90	0.90	0.90	(0.67)	0.50	0.51	0.50	(0.72)
Green tariff, in %	0.02	0.02	0.02	0.02	(0.27)	0.47	0.46	0.46	(0.73)
Heating electricity tariff, in %	0.08	0.08	0.08	0.08	(0.94)	0.03	0.03	0.04	(0.43)
Characteristics at the 1 km^2 grid-level									
Pop. density, in 1k per km^2	1.065	1.071	1.045	1.069	(0.39)	0.591	0.599	0.591	(0.90)
Unemployment rate, in %	5.4	5.3	5.4	5.3	(0.51)	7.8	7.9	7.8	(0.10)
Retirees, in %	21.0	21.0	21.0	21.0	(0.84)	20.7	20.8	20.8	(0.80)
Purch. power, in 1k € per hh	43.6	43.5	43.5	43.6	(0.74)	36.0	36.1	36.0	(0.60)
Foreign household heads, in %	4.2	4.2	4.2	4.2	(1.00)	1.5	1.5	1.5	(0.76)
Number of participants	76,252	12,869	12,841	12,856	$\sum$ = 114,818	4559	1943	1944	$\sum$ = 8446

Note: The columns give averages for participants in the control, monetary framing, ecological framing, and combined framing group. Due to the smaller sample size, we did not implement the combined framing treatment in the REG sample. P-values are from F-tests on mean equality in all experimental groups of SREG and REG, respectively. All regional variables are from 2012. Purchasing power corresponds to the average annual purchasing power of households, and the share of foreign household heads corresponds to the share of non-German household heads, both at the 1 km^2 grid-level.

so-called "classical" measurement error can lead to attenuation bias. However, using group-averages as a proxy variable is an example of a "non-classical" measurement error. While reducing the precision of our estimates, it does not affect their unbiasedness and consistency (Hyslop and Imbens, 2001). In Germany, household-level data at a higher granularity than the 1 km grid-level is typically also unavailable for utilities.

2.3. Pre-specification

Our experimental design and hypotheses were not pre-registered prior to study implementation. Therefore, this section outlines the leeway that we had during data analysis and the boundaries determined by our study design. In our main analyses, we estimate and report effects for all treatment arms and the entire experimental sample (Table 2). These analyses follow directly from our experimental design. Moreover, we differentiate treatment effects between the two utilities and between the two years in our experiment. This differentiation reflects study features that we determined before having access to any outcome data. In particular, we decided to collect multiple years of data and to invite several utilities to take part in the experiment. Our aim to estimate utility-specific treatment effects is also mirrored in the choice to stratify the randomization by utility. For these reasons, we regard the hypothesis tests conducted in Section 3.1 as de-facto pre-specified.

Beyond that, we conduct further exploratory heterogeneity analyses. The first analysis, presented in Table 3, is based on the limited information regarding customers' baseline consumption and tariff, which we received from the participating utilities. Based on this data, we do not see any leeway for data mining in terms of variable selection. Yet, as we compare the treatment effects across (self-defined) subgroups of households with a baseline electricity consumption above the median, the highest quartile, and the highest decile, we prominently label this analysis as exploratory in Section 3.2. Only in an additional step, we use secondary data on socio-demographics at the 1 km grid-level and conduct purely exploratory heterogeneity analyses (Table 4). Whenever our analyses are exploratory, we discuss general patterns that emerge rather than highlighting individual estimates.

We also pre-specified the general conceptual framework of our study in the project proposal, which we submitted before implementing the study. In the proposal, we outline the general research question ("What is the impact of information on residential electricity consumption?"). Furthermore, the proposal is explicit about our intention to conduct the RCT at various sites. Yet, we also made several adjustments to the experimental design after submitting the proposal. For example, our goal to test for framing effects was not included initially, but added during the preparatory phase of our experiment. We also responded to feedback from the qualitative pre-test conducted by *Rheingold* and introduced the third treatment arm that combines our monetary and ecological framing ("Combined"). The proposal is available at the AEA registry (AEARCTR-0006724). It was uploaded after the implementation of the experiment and the finalization of our data analysis.

2.4. Descriptive statistics and balancing

As Table 1 shows, our randomization achieved good covariate balance. In particular, we cannot detect any statistical differences between the control group and the treatment groups (see Columns 5 and 9). As a consequence of different customer bases of SREG and REG, observable characteristics differ substantially between these two utilities. On average, SREG households consume around 9 kWh per day, which corresponds exactly to the average of German households, but substantially exceeds the 7.8 kWh consumed by the average REG household.[5] The percentage of participants who have chosen a green electricity tariff is substantially larger at

[5] From an international perspective, both daily consumption numbers are far less than the average U.S. household of around 34 kWh per day (WEC, 2016), for example, yet match the OECD average relatively well (cf. Andor et al., 2020).

REG. Furthermore, the regional characteristics show that REG households typically live in less densely populated neighborhoods, with larger unemployment rates and a lower average purchasing power, as well as with lower shares of foreign household heads. In Table A11 in the Appendix, we show that the percentage of participants that we cannot observe for the entire time span is low and indistinguishable for the experimental groups, which supports our finding that attrition is not an issue in our study.

3. Treatment effects of information letters

3.1. Average treatment effects

We estimate the Average Treatment Effect (ATE) of the information letters (IL) on electricity consumption using the following differences-in-differences model:

$$Y^n_{i,t} = \alpha_i + \beta_t + \sum_F \omega_F \; IL^F_i \; Post_t + \epsilon_i \tag{1}$$

where $Y^n_{i,t} = Y_{i,t}/Y^c_{i,2014}$ denotes the average daily electricity consumption of household i in billing period t ($Y_{i,t}$), normalized by the average control group consumption in the baseline period 2014 ($Y^c_{i,2014}$). Furthermore, α_i and β_i correspond to household i and billing period t fixed effects, $t \in \{2014, 2015, 2016\}$. IL^F_i denotes a treatment group dummy that equals one if household i receives letters with framing F, where $F \in \{\text{monetary}, \text{ecological}, \text{combined}\}$, and zero otherwise. Furthermore, $Post_t$ is a dummy variable that equals one for the two billing periods 2015 and 2016 after the IL have been sent, while $\epsilon_{i,t}$ designates an idiosyncratic error term. Throughout the analyses, we cluster standard errors at the household level.

In our main analysis, we focus on the conservation effect of our three treatment groups on electricity consumption in 2015 and 2016. We also estimate treatment effects separately by treatment group and year. For this purpose, we construct one treatment dummy IL_i that equals one for all three treatment groups, irrespective of the framing. We then explore how heterogeneity in treatment effects relates to household-specific and neighborhood characteristics. Regarding household-specific characteristics, we test whether effect sizes differ by households' tariff or baseline consumption, as suggested by previous studies on resource conservation (e.g. Ferraro and Price, 2013). We also explore heterogeneity in terms of neighborhood characteristics. For this purpose, we merge our data with the *microm* data set described in the previous section.

Table 2 presents the ATE estimates of the IL. Despite their identical content, the effectiveness of the letters differs considerably between REG and SREG customers. Column 1 shows that the point estimate for the ATE at SREG is very close to zero and not statistically significant. The large sample size enables us to estimate a narrow 95% confidence interval that ranges from −0.33 to 0.18%. The low standard errors from Column 1 of Table 2 translate into a mean detectable effect size of −0.36% (at the conventional 80% power level and 5% level of statistical significance). Put differently, we cannot reject the null hypothesis of no effect for SREG because no sizeable effect exists.

Furthermore, Column 4 shows that REG customers reduce electricity consumption by −1.2% compared to baseline consumption over the two year observation period. This ATE is statistically significant at the five percent level and translates into an absolute reduction of around 36 kWh per year or 0.10 kWh per day, which corresponds to switching off a 30 Watt light bulb for about four hours every day. For comparison, the estimated ATE of −1.2% is similar in magnitude to the ATEs of −0.5% to −3.3% estimated for social comparison-based home energy reports that have received much attention in the literature (e.g. Allcott, 2011, 2015; Andor et al., 2020).[6]

In a next step, we investigate the framing effects of reporting electricity savings in monetary terms (*monetary*), in CO_2 terms (*ecological*), or as a combination of both (*combined*). The point estimates presented in Columns (2) and (5) indicate that for both utilities, the ATE in the ecological framing condition is about twice as large as in the monetary framing condition, reaching −0.18% and −1.6% for SREG and REG, respectively. While we cannot reject the null hypothesis of no difference between the framing conditions at any conventional level, we can reject the null hypothesis that ecologically-motivated savings are ineffective at REG with 95% confidence. Taken together, this suggests that environmentally motivated letters are equally or even more effective than those appealing to monetary motivations. This result is consistent with the findings by Bolderdijk et al. (2013), for example, who show that car owners respond more strongly to environmentally-motivated appeals than to economic appeals.[7]

To investigate whether households respond more strongly in the year of the treatment, we exploit the two billing periods 2015 and 2016 separately. As can be seen in Column 3 of Table 2, we estimate a precise null effect for SREG customers in both years. For REG customers (Column 6), the conservation effect in the year of the treatment amounts to −1.36%, and persists in the year after the treatment has ended, reducing to −1.07%. This decrease implies an annual attenuation rate of about 27%. It is comparable to the persistence of social comparison based home energy reports and similar interventions that have found attenuation rates of around 15 – 50% in the year after the treatment (Allcott and Rogers, 2014; Bernedo et al., 2014; Brandon et al., 2017; Ferraro et al., 2011).

[6] For REG, the mean detectable effect size amounts to −1.43%. While it is larger than for SREG, our study is sufficiently powered to detect typical effect sizes of about −2% that have been found for home energy reports (e.g., Allcott, 2011, 2015).

[7] The fact that the point estimate for the combined treatment group is even slightly positive may call into question the recommendation from the qualitative pretest. Maybe other forms of pretests such as quantitative studies in online samples would have had higher predictive power for the outcomes of this study.

Table 2
ATE by utility, framing condition, and year (in %).

	(1)	(2)	(3)	(4)	(5)	(6)
	SREG			REG		
IL	−0.072			−1.225**		
	(0.130)			(0.512)		
$IL^{monetary}$		−0.069			−0.815	
		(0.194)			(0.623)	
$IL^{ecological}$		−0.181			−1.633**	
		(0.199)			(0.648)	
$IL^{combined}$		0.033				
		(0.201)				
IL × 2015			−0.061			−1.361***
			(0.123)			(0.497)
IL × 2016			−0.085			−1.073*
			(0.165)			(0.625)
R^2	0.014	0.014	0.014	0.003	0.003	0.003
Number of obs.	316,571	316,571	316,571	23,294	23,294	23,294
Number of participants	113,903	113,903	113,903	8359	8359	8359

Note: Standard errors are in parentheses and clustered at the household level. $IL^{monetary}$, $IL^{ecological}$ and $IL^{combined}$ denote the ATEs in our three framing groups, respectively. Due to the smaller sample size, we did not implement the combined framing treatment, $IL^{combined}$, in the REG sample. $IL \times 2015$ and $IL \times 2016$ denote the ATE in the year 2015 and 2016, respectively. In Columns (2), (3), (5), and (6), we estimate Differences-in-Difference-in-Difference models and omit a reference group, so that the estimates correspond to the ATE in the respective subgroup. ***, **,* denote statistical significance at the 1%, 5% and 10% level, respectively.

Table 3
ATE by baseline consumption and tariff (in %).

	(1)	(2)	(3)	(4)	(5)	(6)
Subgroup	SREG			REG		
	ATE	Std. Err.	n	ATE	Std. Err.	n
Baseline cons. ≤ median	−0.162*	(0.086)	157,817	−0.314	(0.379)	11,599
Baseline cons. > median	0.012	(0.243)	158,754	−2.066**	(0.946)	11,695
Baseline cons. > p75	0.224	(0.447)	79,002	−3.629**	(1.743)	5,831
Baseline cons. > p90	0.647	(0.987)	31,330	−4.282	(3.646)	2,316
Green tariff	1.185	(0.961)	5,119	−0.096	(0.599)	10,981
Default tariff	−0.163	(0.117)	294,907	−1.425**	(0.684)	11,493
Heating tariff	0.256	(1.309)	16,545	−14.609**	(6.947)	820

Note: Standard errors are clustered at the household level, standard errors in parentheses. ATEs are estimated in the specified subgroup as described in Eq. (1). ***, **,* denote statistical significance at the 1%, 5% and 10% level, respectively. Participants in the above median, top quartile, and top decile groups consume more than 11.3, 14.5, and 19.2 kWh per day (REG) and 13.4, 17.8, and 24.5 kWh (SREG), respectively.

3.2. Exploratory heterogeneity analysis

The stark contrast in effect sizes between SREG and REG ($IL_{SREG} - IL_{REG}$: −1.153, p-value: 0.029) highlights the importance of understanding response heterogeneity in more detail. We start by exploring how treatment effects relate to households' observable characteristics. Results in this section should be interpreted with care, since we did not pre-specify these subgroup analyses. Also, the sample sizes become small in some of the subgroups, in particular for REG.

We first analyze whether or not the households with higher energy consumption levels realize larger savings, which has been found in prior studies on water (Ferraro et al., 2011; Ferraro and Price, 2013) and energy consumption (Allcott, 2011; Andor et al., 2020). For this purpose, we estimate the ATE for four subgroups of households that use less than the median, more than the median, more than the top quartile, and more than the top decile of baseline electricity consumption, respectively. For REG, the results confirm that customers with higher consumption levels conserve more electricity (Column 4 of Table 3). While households with an electricity consumption below the median realize only a statistically insignificant conservation effect of −0.3%, we detect a statistically significant reduction of −2.1% for households above the median. The treatment effect even reaches −3.6% and −4.3% for households in the top quartile and top decile, respectively. Yet, as the relatively low sample sizes in each of the subgroups give rise to power concerns, we interpret the differences in the point estimates only as suggestive evidence. For SREG, we cannot detect that high-consumption households conserve more (Column 1). For households with consumption levels below the median, we detect a conservation effect of −0.16%, which is small in size but statistically significant at the 10% level. Yet, we cannot detect any electricity savings for households with higher consumption levels.

Next, we test for treatment effect heterogeneity by customers' tariff. Column (4) in Table 3 shows that we detect the largest behavioral response for REG customers with a heating tariff, who reduce electricity consumption by almost −15%. This effect is

Table 4
ATE by neighborhood characteristics (SREG).

	(1)	(2)	(3)	(4)	(5)
IL	−0.070	−0.070	−0.074	−0.074	−0.075
	(0.130)	(0.130)	(0.130)	(0.130)	(0.130)
IL × Density	0.206***				
	(0.061)				
IL × Unemployed		0.096***			
		(0.024)			
IL × Retirees			0.006		
			(0.029)		
IL × PurchPower				−0.039***	
				(0.012)	
IL × HeadForeign					−0.037
					(0.028)
R^2	0.014	0.014	0.014	0.014	0.014
Number of obs.	316,566	316,566	316,390	316,566	316,000
Number of participants	113,901	113,901	113,838	113,901	113,697

Note: Standard errors are clustered at the household level, standard errors in parentheses. Outcome variables are demeaned, so that the parameter estimates on *IL* corresponds to the ATE at the mean. All interaction terms show the change in the ATE (in percentage points per unit of the respective covariate). ***, **,* denote statistical significance at the 1%, 5% and 10% level, respectively.

statistically significant at the 5% level and differs strongly from the ATE for default tariff customers (difference: −13.1 percentage points, p-value: 0.058), and even more from the ATE for green tariff customers (difference: −14.5 percentage points, p-value: 0.037). Heating tariff users are characterized by high consumption levels and thus this finding is in line with the larger effect for high-consumption households. Yet, we caution against overinterpreting the magnitude of this point estimate, given the small sample size of only 820 customers in this subgroup. The estimated conservation effect for default tariff users is at around −1.4% and statistically significant. Again, the results for SREG are different and we cannot detect a strong conservation effect for SREG heating tariff customers (Column 1).

We also explore how socio-demographic neighborhood characteristics are related to the size of treatment effects and estimate Eq. (1) separately for each characteristic. Since such heterogeneity analyses are demanding in terms of statistical power, we only discuss the results for our large sample of SREG customers in the main text. The results for REG are similar, yet less precisely estimated, and can be found in Appendix Table A10.

As the first column of Table 4 shows, the electricity savings of SREG households are less pronounced in densely populated areas. In particular, the ATE increases by 0.2 percentage points as population density increases by 1000 inhabitants per square kilometer. This effect is statistically significant and could potentially explain why SREG customers have lower ATEs, compared to REG customers, who predominantly live in rural areas that are less densely populated. Column (2) shows that the electricity savings are smaller in absolute terms in neighborhoods with large unemployment rates, while households in neighborhoods with larger average purchasing power save more (Column 4).

3.3. Heterogeneity analysis based on causal forests

The explorative heterogeneity analysis based on linear regressions gives a first indication of treatment effect heterogeneity. Yet, detecting heterogeneity patterns based on linear regression analysis is difficult. For example, treatment effect heterogeneity is often non-linear in covariates and may depend on complex interaction effects. When researchers select which variables and interaction effects to include based on estimation results, this procedure may lead to a selection of statistically significant, but spurious, heterogeneity patterns (see, e.g., Athey and Imbens, 2016 for a general discussion of "honest" estimation). To overcome these limitations, we explore treatment effect heterogeneity based on recently developed causal forest methods (Wager and Athey, 2018; Athey et al., 2019). This methodology allows us to assess how treatment effect heterogeneity relates to observable characteristics. It also allows us to test the robustness of the heterogeneity patterns that we have identified in our previous heterogeneity analysis (Section 3.2).

Causal forests adapt machine learning algorithms to the estimation of treatment effects in large samples. This algorithm estimates conditional average treatment effects CATE(x), i.e., the ATE at particular covariate realizations $x \in X$, where X denotes the covariate space. In our case, the covariate space corresponds to all attribute combinations of the household-specific and neighborhood characteristics. In particular, we consider the following covariates in our setting: the baseline electricity use, dummies for electricity tariffs, the population density, the unemployment rate, the average purchasing power per household, the percentage of retirees, and the percentage of foreign household heads.

The basic building block of a causal forest algorithm is a "causal tree", which is constructed based on a randomly selected subsample of the data, the so-called "root node". For identification, a causal tree partitions the covariate space into subsamples with similar CATEs. Partitioning minimizes a mean squared error criterion for treatment effects in order to maximize treatment effect prediction accuracy. To avoid overfitting, partitioning penalizes treatment-control imbalance and the variance in ATEs within a node. Once further splits of the data do not increase the error criterion, a final partition is reached, the so-called "leaves".

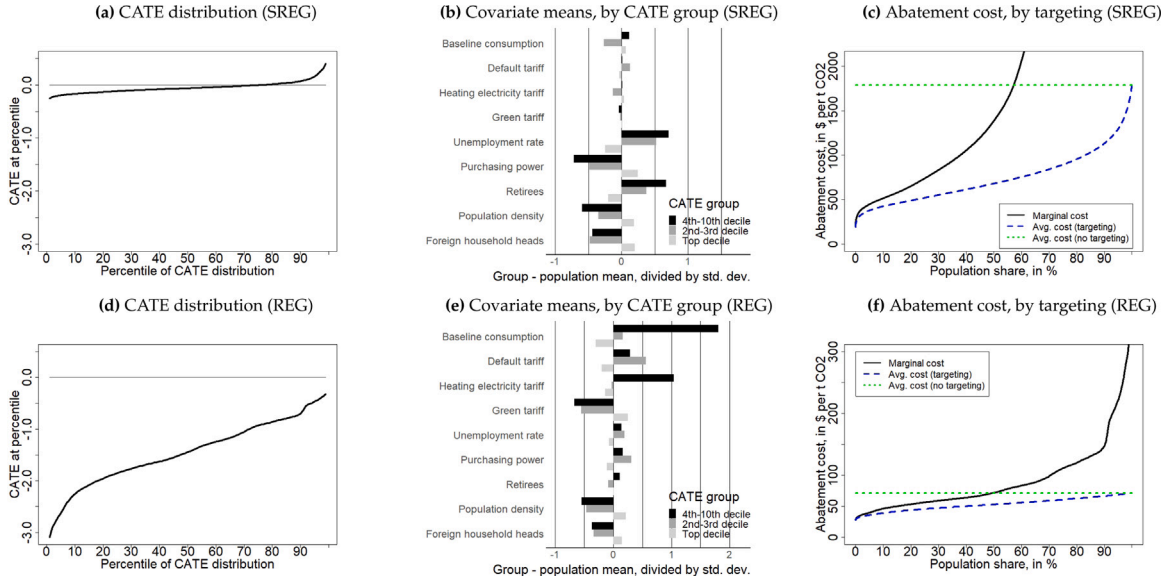

Fig. 2. Targeting using causal forest methods for estimating CATE. *Note:* We estimate CATE based on a causal forest (Wager and Athey, 2018). We follow Athey et al.'s (2019) recommendation to grow a large number of trees (4000) and to determine tuning parameters based on a cross-validation procedure. Details on the estimation and the tuning parameters can be found in Appendix A6. Figs. 2(a) and 2(d) give the resulting CATE estimates, sorted by CATE percentile for SREG and REG, respectively. Figs. 2(b) and 2(b) give the difference between the characteristics of households in the top decile, the second to fourth decile, and above the fourth decile, compared to the overall sample, divided by its standard deviation. Figs. 2(f) and 2(f) present the abatement cost of households at a particular percentile of the CATE distribution ("Marginal cost"), the average abatement cost of sending letters to all households whose CATE is smaller or equal to the CATE at a particular percentile ("Avg. cost (targeting)"), and the average abatement cost of sending the letter to all customers ("Avg. cost (no targeting)"). To calculate abatement cost, we approximate intervention cost with 4$ (1$ per letter), use the average German carbon intensity of 486 g per kilowatt-hour (IEA, 2015), neglect discount rates, and assume that treatment effects decrease linearly by around 20 percentage points per annum, as implied by our estimates for REG households.

For every leave, the CATE is then estimated based on observations from another subsample. This so-called "honest" approach (Athey and Imbens, 2016) ensures consistency of the CATE estimator despite the fact that the partitions are determined in a data-driven manner. A causal forest algorithm repeats this process for other randomly selected root nodes and averages the tree-specific CATEs. To deal with the panel nature of our data set, we transform the outcome variable to first differences ($Y^n_{i,2015} - Y^n_{i,2014}$) and thus use the causal forest algorithm to analyze heterogeneity in the treatment period 2015. As the estimation of a random forest can be sensitive to the choice of some tuning parameters, such as the minimum number of observations per leaf, we follow (Athey et al., 2019) and determine those parameters optimally through cross-validation (for details, see Appendix A6).

We first assess the extent of treatment effect heterogeneity by plotting conditional average treatment effect (CATE) estimates for SREG and REG households, respectively, against effect size percentiles (Figs. 2(a) and 2(d)). Fig. 2(a) shows that the CATEs of SREG households are small overall. For of the most responsive SREG households, i.e., those at the first percentile of the effect size distribution, it amounts to about −0.3%, which is only slightly less than the ATE of −0.072% (Section 3.1). Furthermore, we estimate a negative CATE for the vast majority of households. Positive CATE estimates are rare and always smaller than 0.4%. This finding implies that the low ATE for SREG customers stems from low effect sizes overall rather than from heterogeneous positive and negative CATEs that cancel out on average. As shown in Fig. 2(d), we detect substantially more treatment effect heterogeneity for REG. About 15% of the households have treatment effects beyond −2% and about 60% of all households reduce their electricity consumption by more than −1%. Furthermore, CATE estimates are exclusively negative.

In Figs. 2(b) and 2(e), we explore the link between observable household characteristics and the magnitude of the CATE estimates. For that purpose, we define three groups of households. A first group consists of households in the top CATE decile (i.e., the 10% of households with the largest electricity savings), a second group consists of households in the second and third CATE deciles, and a third group consists of all remaining households. For every covariate, we calculate the difference between the covariate mean in a given group and the respective population mean, normalized by the standard deviation of that covariate.

Based on these standardized differences, we investigate how responsive households, for example those in the first CATE decile, differ from the average household in terms of covariates. For both utilities, we find that responsive households live in less densely populated areas and are less likely to have foreign household heads, which mirrors the results from our univariate heterogeneity analyses. Beyond this similarity, the heterogeneity patterns differ. For REG, we find that households in the top decile of treatment effects have larger baseline electricity consumption and are more likely to have a heating tariff compared to the overall population, with a normalized difference of about 1.8 and 1.0 standard deviations for REG households, respectively. For SREG customers, this pattern does not hold. Responsive SREG households tend to live in neighborhoods with higher unemployment rates and lower purchasing power.

Table 5
The effect of targeting on the cost-effectiveness of IL.

		(1) IL recipients, in %	(2) ATE of IL recip., in %	(3) Avg. abatement cost, in $ per t
(a) No targeting				
SREG		100	−0.06	1788.6
REG		100	−1.4	71.3
(b) Extrapolations across utilities, based on matching				
Matched SREG sample (with REG characteristics)		100	−0.28	1062.1
(c) Extrapolations across utilities, based on heterogeneity patterns (CATEs) and covariates				
Extrapolation from SREG to REG: $E_X\left[CATE^{SREG}(X^{REG})\right]$		100	−0.12	842.2
Extrapolation from REG to SREG: $E_X\left[CATE^{REG}(X^{SREG})\right]$		100	−1.9	45.2
(d) Utility-specific targeting under the following policy objectives				
SREG	Max. abatement	77	−0.1	911.7
	Max. benefit (SCC: 119$ per t CO_2)	0	–	–
	Max. benefit (SCC: 41$ per t CO_2)	0	–	–
	Max. benefit (SCC: 12$ per t CO_2)	0	–	–
REG	Max. abatement	100	−1.4	71.3
	Max. benefit (SCC: 119$ per t CO_2)	79	−1.7	62.2
	Max. benefit (SCC: 41$ per t CO_2)	6	−2.8	36.3
	Max. benefit (SCC: 12$ per t CO_2)	0	–	–

Note: Our calculations are based on the CATE estimates for SREG and REG from Figs. 2(a) and 2(d), respectively. To calculate abatement cost, we approximate intervention cost with 4$ (1$ per letter), use the average German carbon intensity of 486 g per kilowatt-hour (IEA, 2015), neglect discount rates, and assume that treatment effects decrease linearly by around 20 percentage points per annum, as implied by our estimates for REG households. *ATE of IL recipients* denotes the average treatment effect and *Avg. abatement cost* denotes the average abatement cost for information letter (IL) recipients, respectively. We consider three targeting schemes: *No targeting* implies that all households receive IL, *Max. abatement* targets households whose predicted treatment effects exceed zero, and *Max. benefit* targets households whose letter cost per saved ton of CO_2 is lower than the following three assumed social cost of carbon (SSC) of 119, 41, and 12$ per t CO_2, respectively. *Extrapolations Across Utilities, Based on Matching*, conducts 1-to-1 propensity score matching (without replacement) using all regional variables and then estimates the ATE in the matched SREG sample. *Extrapolation from SREG to REG*, for example, estimates the ATE at REG, using the CATE estimates from SREG and the covariates of REG customers.

4. Cost-effectiveness, external validity, and targeting

In this section, we assess the cost-effectiveness of our intervention, the ability to extrapolate findings from one utility to another, and the potential of targeting. For every participant, we calculate the implied abatement cost per ton CO_2, which is a widely-used benchmark for assessing the cost of energy-saving measures.[8] We compare these costs with the benefits from the avoided social cost of carbon (SCC). As the size of the SCC is subject to dispute in the literature, we use an estimate of 41$ per ton of CO_2, but also consider 12 and 119$ per ton of CO_2 as lower and upper bound estimates (IAWG, 2016, all values deflated to 2015 $).

First, we assess the cost-effectiveness of sending the information letters to all households at the SREG and REG in our sample. As shown by Table 5, the average abatement cost for SREG customers amount to 1789$ per ton of CO_2, which clearly exceeds even a large SCC estimate of 119$. Hence, the benefits of a climate policy that sent information letters to all SREG customers would fall below its costs. For REG customers, we find that the average abatement cost amount to 71$, which is below our upper bound SCC estimate of 119$. A policy of sending information letters to these customers can thus be rationalized if policy makers expect rather large damages from global warming.

Next, we explore the implications of the large differences in effect sizes and average abatement cost on the ability to extrapolate treatment effect across study populations. For example, if we had only run our experiment at REG and had tried to quantitatively extrapolate our results to SREG, would we have been able to predict the average treatment effect at that utility? Clearly, using our ATE estimate of −1.4% from REG as a predictor of the ATE at SREG would have resulted in strong prediction errors. One reason for such errors is that a naive prediction does not take differences in sample characteristics into account. To test whether or not differences in characteristics can explain the differences across utilities, we employ two strategies. First, we use matching to find observations within the SREG sample that match those of the REG sample in terms of all regional covariates. Second, we use our CATE estimates from one utility and predict the ATE for customers of the other utility.

As shown in Panel (b) of Table 5, we do not find that differences in observables explain the difference in ATEs across utilities. When we use one-to-one propensity score matching to construct a sample of those SREG customers that match REG customers in terms of all regional variables, we estimate an average treatment effect of −0.28%, which is still considerably lower than the −1.4% found for REG customers. A similar picture emerges when we use our causal forest estimates to extrapolate across utilities (Panel

[8] We use our participant-specific CATE estimates, approximate intervention cost with 1$ per letter, use the average German carbon intensity of 486 g per kilowatt-hour (IEA, 2015), neglect discount rates, and assume that treatment effects decrease linearly by around 20 percentage points per annum.

c).[9] When we predict the ATE in the SREG sample based on the CATEs estimated for REG customers, we obtain an ATE of −1.9%, which is even larger than the ATE for REG customers.[10] This large effect translates into a predicted average abatement cost of 45$ per ton CO_2, which is considerably below an SSC estimate of 119$, for example. Furthermore, predicting the ATE for REG customers based on the CATEs estimated for SREG customers yields an ATE of only −0.12% and average abatement cost of more than 800$.

This finding demonstrates that using socio-demographic data to extrapolate results across utilities would lead to large prediction errors and misguided policy recommendations. It holds true despite the fact that the information letters had identical content, that the program implementation was indistinguishable across utilities, and that we use comprehensive data on socio-economic characteristics and a sophisticated causal forest machine learning algorithm to account for differences in sample composition. The large unexplained heterogeneity in treatment effects points to the importance of utility-specific moderating factors. To give an example, different customer engagement habits and differences in reputation across utilities may co-determine whether customers actively read the letters and trust the information that is provided. In addition, customers may select into a utility based on unobservable characteristics, which could also moderate the impact of the intervention. Such contextual factors are difficult to quantify and pose substantial challenges for generalizing heterogeneity patterns and effect sizes across utilities.

Differences in context also arise from the fact that customers of both utilities tend to live in different geographic areas (see Appendix Figure A10 for an overview). SREG recruits its customers from all over Germany (with a concentration in the center and south-east of the country), including large cities and the densely populated agglomerations. By contrast, REG supplies virtually only customers in the very north-eastern part of the country. Beyond what standard socio-demographic data can capture, regional differences may correlate with unobservable characteristics of consumers and may hence moderate treatment effects. In addition, households in some different geographic areas may have been exposed to earlier energy efficiency awareness campaigns that were conducted by some local utilities in the 1970s in response to the oil crisis.[11]

In Panel (d) of Table 5, we explore the extent to which targeting based on utility-specific heterogeneity patterns could increase the effectiveness of interventions at the same utility. As Figs. 2(c) and 2(f) illustrate, targeting can substantially reduce the average abatement cost from the intervention when information letters are sent only to all households with the most pronounced CATEs and, hence, the lowest abatement cost (denoted as marginal abatement cost in Figs. 2(c) and 2(f)).

Next, we quantify the benefits of two targeting strategies. First, a policy maker may want to maximize total CO_2 abatement by sending letters to all households with negative CATE estimates. Second, a policy maker may aim to maximize the net benefits of the intervention, defined as the difference between the avoided social cost of carbon and the abatement cost. For this targeting strategy, we again consider the three different scenarios based on a social cost of carbon of 12, 41, and 119$ per ton, respectively.

For SREG customers, we find that targeting does not help to increase the effectiveness of the intervention. Maximizing abatement by sending information letters only to the 77% of customers with negative CATE would only marginally increase average energy savings to −0.1%. This finding reflects the absence of significant treatment effect heterogeneity at SREG (see Fig. 2(a)). Furthermore, a targeting strategy that aims to maximize social benefits of the intervention would not send information letters to any SREG household, no matter which social cost of carbon estimate we consider. For REG customers, we find that a strategy to achieve maximum abatement would send information letters to all households which would result in abatement cost of about 71$. A targeting strategy that maximizes social benefits at social costs of carbon of 119 and 41$, would send letters to 79 and 6%, respectively, at average abatement costs of 62 and 36$, respectively. Hence, at REG, targeting could substantially reduce average abatement cost, while sending letters to a sizeable percentage of households. Only when assuming a low social cost of carbon of 12$ do we find that information letters would be sent to none of the REG customers.

5. Conclusion

Based on a large-scale randomized controlled trial among more than 120,000 customers of two utilities, this paper has evaluated the effectiveness of a letter-based information campaign about tips for energy efficient investments and behaviors. In a well-powered experiment, we find that the average effect size is small, irrespective of the framing. By contrast, we detect substantial heterogeneity in treatment effects between both utilities. While we estimate a precise null effect (−0.06%) in the year of the treatment for customers of the larger supra-regional utility (SREG), consumers of the smaller regional utility (REG) reduce their electricity consumption much more strongly, by −1.4%. This effect at REG persists one year after the treatment has ended, but decreases by 27 percentage points.

Beyond estimating the average treatment effects, we conduct explorative heterogeneity analyses to identify particularly responsive consumer subgroups. In line with previous energy conservation studies (Allcott, 2011; Andor et al., 2020), we find that REG customers with high baseline electricity consumption levels exhibit higher effect sizes than those with low consumption levels. For SREG we do not find these differences, though. To explain heterogeneity across utilities, we leverage the full potential of our data set and employ a causal forest machine learning algorithm developed by Wager and Athey (2018). Using comprehensive socio-demographic data, we show that heterogeneity in treatment effects cannot be explained by differences in the observable characteristics of customers. As the letter content and the implementation of the program were identical for both utilities, this finding points to the importance of utility-specific factors, such as differences in unobserved customer characteristics and customer engagement habits.

[9] We present the distributions of estimated and predicted CATEs in Appendix Figure A11.

[10] The reason for larger ATEs is that high-usage households have larger treatment effects at REG, but not at SREG, and that SREG households tend to have larger consumption levels overall.

[11] For the two utilities in our experiment, we can exclude that energy efficiency awareness campaigns were conducted in the years prior to our field test.

The finding that our information letters yield a −1.4% reduction for one utility also adds to the understanding of the effectiveness of home energy reports (HER), which have been proposed as a promising policy tool by which to reduce energy consumption (Allcott and Mullainathan, 2010). The literature on HER has typically attributed their conservation effect of about −0.5% to −3.3% to the presence of a social comparison module, despite the fact that HER typically also include other elements, such as energy saving tips (e.g. Allcott, 2011, 2015). Our finding shows that a social comparison might not be the only element of such letters that triggers energy conservation. In fact, social comparison based home energy reports have been shown to realize about −0.7% in Germany (Andor et al., 2020); only about half of the conservation effect that our information letters achieve at REG.

Our findings have important implications for policy. In contrast to previous studies on retirement savings (Dolls et al., 2018) and social comparison based reports (e.g. Allcott, 2011, 2015), our evidence suggests that letter-based information campaigns are largely ineffective when used as a universal policy. While we detect sizeable effect sizes for particular consumer subgroups, we find that these groups cannot be identified based on observable characteristics. The presence of site-specific factors represents a significant obstacle for bringing an informational intervention to scale. First, they complicate learning from a pilot study about the effect sizes of the same intervention at another site, or even in the overall population. If a pilot had only been conducted in the REG sample, for example, policy makers would have made misinformed scaling decisions by wrongly expecting that these considerable savings would also materialize elsewhere (e.g. in the SREG sample). Second, site-specific factors also prevent the derivation of generally applicable targeting strategies that could otherwise allow the cost-effectiveness of informational interventions to improve.

More broadly, our findings provide further evidence on the extent to which the causal effects measured in a particular study population and set-up depend on the particular context and the implementation partner (e.g. Allcott, 2015; Dehejia et al., 2019; Gechter, 2016; Peters et al., 2018; Vivalt, 2020). In particular, Vivalt (2020) finds that generalizability between different programs and settings is very limited for many types of interventions and recommends to conduct impact evaluations in multiple settings with varying contexts. Our study implements this suggestion and shows that context matters: identical treatments delivered in the same country can induce sizeable savings in some study populations, while being virtually ineffective in others. The fact that our available socio-demographic variables fail to explain the large differences in treatment effects across utilities points to the importance of other factors that may be difficult to quantify. In our context, such factors include the reputation of a utility, which may affect how households perceive information letters. Another potential factor is the exposure to information prior to the intervention, which may be higher in regions where environmental protection organizations, utilities, and schools are more active in disseminating it. Further work that assesses potential site differences ex-ante and helps to disentangle the importance of such partner- and site-specific moderators would be valuable. In particular, it could allow to better evaluate the population-wide effects of a wide range of interventions that involve energy utilities, hospitals, or schools.

Acknowledgments

We gratefully acknowledge financial support from the Stiftung Mercator. Furthermore, this work has been partly supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under CRC TR 224 and the Collaborative Research Center *Statistical Modeling of Nonlinear Dynamic Processes* (SFB 823), within Project A3, *Dynamic Technology Modeling*. We are grateful for valuable comments and suggestions by Nathan Fiala, Kenneth Gillingham, Lorenz Götte, Timo Goeschl, Katrina Jessoe, Andreas Lange, Andreas Löschel, Grischa Perino, and Mike Price, as well as participants at the 33rd Annual Congress of the European Economic Association, the 6th World Congress of Environmental and Resource Economists, the 1st RWI Empirical Environmental Economics Workshop, the ASSA 2020 Annual Meeting, the 2020 AERE Summer Conference, as well as seminars at Yale, Vrije Universiteit Amsterdam, University of Hamburg, University of Münster, and ZEW. We thank E.ON SE and WEMAG AG, in particular Melanie Lemke, Michael Paul, Kristina Rodig, and Sonja Sellnau, for their cooperation in implementing the study and Lukas Tomberg for excellent research assistance.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.euroecorev.2022.104094.

References

Al-Ubaydli, O., List, J.A., LoRe, D., Suskind, D., 2017a. Scaling for economists: Lessons from the non-adherence problem in the medical literature. J. Econ. Perspect. 31 (4), 125–144.

Al-Ubaydli, O., List, J.A., Suskind, D.L., 2017b. What can we learn from experiments? Understanding the threats to the scalability of experimental results. Am. Econ. Rev. 107 (5), 282–286.

Allcott, H., 2011. Social norms and energy conservation. J. Public Econ. 95 (9), 1082–1095.

Allcott, H., 2015. Site selection bias in program evaluation. Q. J. Econ. 130 (3), 1117–1165.

Allcott, H., Kessler, J.B., 2019. The welfare effects of nudges: A case study of energy use social comparisons. Am. Econ. J. Appl. Econ. 11 (1), 236–276.

Allcott, H., Knittel, C., 2017. Are Consumers Poorly Informed about Fuel Economy? Evidence from Two Experiments. NBER Working Papers 23076, National Bureau of Economic Research.

Allcott, H., Mullainathan, S., 2010. Behavior and energy policy. Science 327 (5970), 1204–1205.

Allcott, H., Rogers, T., 2014. The short-run and long-run effects of behavioral interventions: experimental evidence from energy conservation. Am. Econ. Rev. 104 (10), 3003–3037.

Allcott, H., Sweeney, R.L., 2017. The role of sales agents in information disclosure: Evidence from a field experiment. Manage. Sci. 63 (1), 21–39.

Andor, M.A., Fels, K.M., 2018. Behavioral economics and energy conservation–A systematic review of non-price interventions and their causal effects. Ecol. Econ. 148, 178–210.

Andor, M.A., Gerster, A., Peters, J., Schmidt, C.M., 2020. Social norms and energy conservation beyond the US. J. Environ. Econ. Manage. 103, 102351.
Asensio, O.I., Delmas, M.A., 2015. Nonprice incentives and energy conservation. Proc. Natl. Acad. Sci. USA 112 (6), E510–E515.
Athey, S., Imbens, G., 2016. Recursive partitioning for heterogeneous causal effects. Proc. Natl. Acad. Sci. USA 113 (27), 7353–7360.
Athey, S., Tibshirani, J., Wager, S., 2019. Generalized random forests. Ann. Statist. 47 (2), 1148–1178.
Bernedo, M., Ferraro, P.J., Price, M., 2014. The persistent impacts of norm-based messaging and their implications for water conservation. J. Consum. Policy 37 (3), 437–452.
Bhargava, S., Manoli, D., 2015. Psychological frictions and the incomplete take-up of social benefits: Evidence from an IRS field experiment. Am. Econ. Rev. 105 (11), 3489–3529.
Bolderdijk, J.W., Steg, L., Geller, E.S., Lehman, P., Postmes, T., 2013. Comparing the effectiveness of monetary versus moral motives in environmental campaigning. Nat. Clim. Change 3 (4), 413.
Brandon, A., Ferraro, P.J., List, J.A., Metcalfe, R.D., Price, M.K., Rundhammer, F., 2017. Do the Effects of Social Nudges Persist? Theory and Evidence from 38 Natural Field Experiments. Technical Report. NBER Working Paper No. 23277.
Buckley, P., 2020. Prices, information and nudges for residential electricity conservation: A meta-analysis. Ecol. Econ. 172, 106635.
Byrne, D.P., Nauze, A.L., Martin, L.A., 2018. Tell me something i don't already know: Informedness and the impact of information programs. Rev. Econ. Statist. 100 (3), 510–527.
Carter, S.P., Skimmyhorn, W., 2018. Can information change personal retirement savings? Evidence from social security benefits statement mailings. In: AEA Papers and Proceedings. 108, pp. 93–97.
Dehejia, R., Pop-Eleches, C., Samii, C., 2019. From local to global: External validity in a fertility natural experiment. J. Bus. Econ. Statist. 0 (0), 1–27.
Delmas, M.A., Fischlein, M., Asensio, O.I., 2013. Information strategies and energy conservation behavior: A meta-analysis of experimental studies from 1975 to 2012. Energy Policy 61, 729–739.
Dolls, M., Doerrenberg, P., Peichl, A., Stichnoth, H., 2018. Do retirement savings increase in response to information about retirement and expected pensions?. J. Public Econ..
Duflo, E., Saez, E., 2003. The role of information and social interactions in retirement plan decisions: Evidence from a randomized experiment. Q. J. Econ. 118 (3), 815–842.
Ferraro, P.J., Miranda, J.J., Price, M.K., 2011. The persistence of treatment effects with norm-based policy instruments: Evidence from a randomized environmental policy experiment. Am. Econ. Rev. 101 (3), 318–322.
Ferraro, P.J., Price, M.K., 2013. Using non-pecuniary strategies to influence behavior: Evidence from a large scale field experiment. Rev. Econ. Statist. 95 (1), 64–73.
Gechter, M., 2016. Generalizing the results from social experiments: theory and evidence from mexico and india. URL: http://www.personal.psu.edu/mdg5396/Gechter_Generalizing_Social_Experiments.pdf (accessed June 6, 2017).
Goette, L., Leong, C., Qian, N., 2019. Motivating household water conservation: A field experiment in singapore. PLoS One 14 (3), 1–15.
Hanna, R., Mullainathan, S., Schwartzstein, J., 2014. Learning through noticing: Theory and evidence from a field experiment. Q. J. Econ. 129 (3), 1311–1353.
Hyslop, D.R., Imbens, G.W., 2001. Bias from classical and other forms of measurement error. J. Bus. Econ. Statist. 19 (4), 475–481.
IAWG, 2016. Technical support document: Technical update of the social cost of carbon for regulatory impact analysis. URL: https://www.epa.gov/sites/production/files/2016-12/documents/sc_co2_tsd_august_2016.pdf U.S. Interagency Working Group on Social Cost of Greenhouse Gases.
IEA, 2015. CO2 Emissions from Fuel Combustion - Highlights. International Energy Agency, URL: https://www.iea.org/publications/freepublications/publication/CO2EmissionsFromFuelCombustionHighlights2015.pdf (accessed October 28, 2016).
Jaime Torres, M.M., Carlsson, F., 2018. Direct and spillover effects of a social information campaign on residential water-savings. J. Environ. Econ. Manage. 92, 222–243.
Jalan, J., Somanathan, E., 2008. The importance of being informed: Experimental evidence on demand for environmental quality. J. Dev. Econ. 87 (1), 14–28.
Kahneman, D., Tversky, A., 1979. Prospect theory: An analysis of decision under risk. Econometrica 47 (2), 263–292.
Karlin, B., Zinger, J.F., Ford, R., 2015. The effects of feedback on energy conservation: A meta-analysis. Psychol. Bull. 141 (6), 1205.
Knittel, C.R., Stolper, S., 2019. Using Machine Learning to Target Treatment: The Case of Household Energy Use. National Bureau of Economic Research (NBER) Working Paper No. 26531.
Levin, I.P., Schneider, S.L., Gaeth, G.J., 1998. All frames are not created equal: A typology and critical analysis of framing effects. Organ. Behav. Hum. Decis. Process. 76 (2), 149–188.
Levitt, S.D., List, J.A., 2009. Field experiments in economics: The past, the present, and the future. Eur. Econ. Rev. 53 (1), 1–18.
Meager, R., 2019. Understanding the average impact of microcredit expansions: A Bayesian hierarchical analysis of seven randomized experiments. Am. Econ. J. Appl. Econ. 11 (1), 57–91.
microm, 2015. Socioeconomic regional data. version: 1. rwi – rheinisch–westfälisches institut für wirtschaftsforschung e.v. http://dx.doi.org/10.7807/microm:einwGeAl:V3, http://dx.doi.org/10.7807/microm:alq:V3, http://dx.doi.org/10.7807/microm:kaufkraft:V3.
Ofosu, G.K., Posner, D., 2019. Pre-analysis plans: A stocktaking.
Ofosu, G.K., Posner, D.N., 2020. Do pre-analysis plans hamper publication? AEA Pap. Proc. 110, 70–74.
Peters, J., Langbein, J., Roberts, G., 2018. Generalization in the tropics – development policy, randomized controlled trials, and external validity. World Bank Res. Obs. 33 (1), 34–64.
Schultz, P.W., Nolan, J.M., Cialdini, R.B., Goldstein, N.J., Griskevicius, V., 2007. The constructive, destructive, and reconstructive power of social norms. Psychol. Sci. 18 (5), 429–434.
Schwartz, D., Fischhoff, B., Krishnamurti, T., Sowell, F., 2013. The hawthorne effect and energy awareness. Proc. Natl. Acad. Sci. USA 110 (38), 15242–15246.
Stigler, G.J., 1961. The economics of information. J. Polit. Econ. 69 (3), 213–225.
Tiefenbeck, V., Goette, L., Degen, K., Tasic, V., Fleisch, E., Lalive, R., Staake, T., 2018. Overcoming salience bias: How real-time feedback fosters resource conservation. Manage. Sci. 64 (3), 983–1476.
Tiefenbeck, V., Staake, T., Roth, K., Sachs, O., 2013. For better or for worse? Empirical evidence of moral licensing in a behavioral energy conservation campaign. Energy Policy 57, 160–171.
Tonke, S., 2020. Imperfect procedural knowledge: A field experiment to encourage water conservation. doi:https://drive.google.com/file/d/1z6n_dge0JPBlGnrVMuXSoZtqeH7REuX0/view.
Tversky, A., Kahneman, D., 1981. The framing of decisions and the psychology of choice. Science 211 (4481), 453–458.
Vivalt, E., 2020. How much can we generalize from impact evaluations? J. Eur. Econ. Assoc. 18, http://dx.doi.org/10.1093/jeea/jvaa019.
Wager, S., Athey, S., 2018. Estimation and inference of heterogeneous treatment effects using random forests. J. Amer. Statist. Assoc. 13, 1–15.
WEC, 2016. Energy Efficiency Indicators via Enerdata, Average Electricity Consumption per Electrified Household. World Energy Council, URL: https://www.wec-indicators.enerdata.eu/secteur.php/household-electricity-use.html (accessed October 28, 2016).
Wichman, C.J., 2017. Information provision and consumer behavior: A natural experiment in billing frequency. J. Public Econ. 152, 13–33.