Journal of Environmental Economics and Management 132 (2025) 103163

Contents lists available at ScienceDirect

Journal of Environmental Economics and Management

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

The impact of real-time energy consumption feedback on residential gas and electricity usage☆

Mirthe Boomsma [a,b], Kees Vringer [b], Daan van Soest [a,*]

[a] Department of Economics and Tilburg Sustainability Center, Tilburg University, The Netherlands
[b] PBL Netherlands Environmental Assessment Agency, The Netherlands

ARTICLE INFO

JEL classification:
Q41
Q58
D12
D83

Keywords:
Residential energy use
Energy savings
Energy consumption feedback
Smart metering
Climate change
Energy transition
Randomized controlled trial

ABSTRACT

We implemented a Randomized Controlled Trial to estimate the impact of real-time energy consumption feedback on residential energy use. A random half of the participant households were offered an energy consumption monitor that uses smart-meter data to relay real-time information on energy use. The feedback resulted in natural gas and electricity savings of, respectively, 6.9 and 2.2 percent; the total amount of energy consumed was reduced by 5.8 percent. We probe the underlying mechanisms by means of a survey and find suggestive evidence that the display's main impact is via raising the salience of the costs of using energy.

1. Introduction

In the world-wide effort to reduce global greenhouse gas emissions, much of the policy makers' attention is directed towards the residential energy market. Not only is the residential sector responsible for a substantial share of CO_2 emissions (about 14% of CO_2 emissions in the OECD; see OECD, 2023), the amount of energy consumed may be too high even from a private perspective. Many energy-saving measures are not implemented even if they are cost-effective (Allcott and Greenstone, 2012; Gillingham and Palmer, 2014). For instance, households tend to underestimate the cost of using energy for a large number of appliances (and especially so for the more energy-intensive uses; see Attari et al., 2010), and they also tend to pay limited attention to energy costs unless these are sufficiently salient (see for example Allcott, 2011a; Sexton, 2015; Pellerano et al., 2017; Deryugina et al., 2020). Interventions that are able to mitigate households' decision biases may thus result in a "behavioral double dividend" – lower CO2 emissions, and higher private welfare (Allcott et al., 2014).

☆ We thank !Woon, the Woonbond, Agem Energieloket, Utrecht municipality, tenant council Albanianae, tenants platform MEVM, housing association UWoon and residents' initiative Energierijk Houten for the very productive collaboration in implementing this study. We are especially grateful to Eef Meijerman, Ingrid Houtepen, Kelly Schwegler, Jaap van Leeuwen, Marion Overberg, Justin Pagden, Jeroen aan het Rot, Belinda Haverkamp, Cees Veerman, Henk Oostland, Pier Schipper, Iris Uittien, Annet Bultman and Katrin Larsen. Furthermore, we thank Ernestine Elkenbracht and Erik van Lidth de Jeude (Quintens) for managing the project's implementation, and Cheyenne Ramada for her excellent research assistance. We thank The Netherlands Enterprise Agency (RVO) and the Dutch Ministry for Internal Affairs (BZK) for financial support of the project. Finally we thank Ben Vollaard and three anonymous reviewers for their very valuable comments on an earlier version of this paper.

* Corresponding author.
E-mail address: d.p.vansoest@tilburguniversity.edu (D.v. Soest).

https://doi.org/10.1016/j.jeem.2025.103163
Received 1 June 2023
Available online 8 May 2025
0095-0696/© 2025 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

From the range of behaviorally-oriented policy measures, energy consumption feedback is among the strategies that have received most attention; for overviews see Karlin et al. (2015), Zangheri et al. (2019), Khanna et al. (2021) and Agarwal et al. (2023). The key idea is that if households are better informed about their actual energy consumption and associated costs, they will start making better decisions regarding both their energy consumption behaviors and the adoption of more energy-efficient appliances and technologies (Jessoe and Rapson, 2014; Allcott and Rogers, 2014). The typical finding is that feedback on the amount of energy consumed reduces subsequent usage by between 2% and 5% (see Degen et al., 2013; McKerracher and Torriti, 2013; Khanna et al., 2021), even though some studies find savings up to 20% (see Abrahamse et al., 2005; Darby, 2006; Ehrhardt-Martinez et al., 2010; Delmas et al., 2013) – possibly also because the impact seems to vary with the medium of feedback provision (Fischer, 2008; Karlin et al., 2015; Henry et al., 2019).

In this paper, we estimate the impact of providing real-time feedback on households' consumption of both electricity and natural gas. We do so by implementing a Randomized Controlled Trial (RCT) involving over 800 households living in seven regions in the Netherlands. We advertised the existence and potential benefits of an energy monitor, an In-Home Display (IHD), that, using smart meter technology,[1] is able to provide information, real-time and separately for natural gas and electricity, about the household's current energy consumption. We announced that households could apply to receive the IHD, free of charge, for a period of one year, and that a lottery would determine which of the applicant households were to receive a device. We specified that applicant households should live in homes using both natural gas and electricity and equipped with a smart meter, and that they should have the intention to continue living in their current home for at least one more year. Importantly, participation in the lottery was also conditional on households' willingness to share their daily energy consumption data independent of whether they ended up winning an IHD, or not. Identification of the treatment effect is thus by means of a random rejection design, in which about half of the applicants were offered a device, and the other half did not.

We find that the display resulted in a 6.9 percent decrease in the use of gas and in a 2.2 percent reduction in the amount of electricity consumed; overall, this translates into aggregate energy savings of 5.8 percent. While the induced savings on gas consumption vary with neither the season nor baseline energy consumption, we find suggestive evidence that the IHD was especially effective in reducing electricity consumption among the high-intensity electricity users. We complement the analysis of the smart meter data with the results of a survey. Survey responses suggest that users' assessment of the device's usefulness did not decline (much) over time, and that a large share of the respondents still use the display long after having received it. Despite the fact that the overall response rate is relatively high (at 55%), (some) self-selection into the survey cannot be ruled out, and hence we take this as suggestive evidence that continued exposure to high-frequency consumption feedback may have been key for sustaining the achieved energy savings – especially because the IHD seems to have enhanced the salience of the costs of using energy.

We contribute to five fields within the wider literature on energy saving and information provision. First, we complement the literature on the effectiveness of energy use feedback (see for example Schultz et al., 2007; Jessoe and Rapson, 2014; Allcott and Rogers, 2014; Harold et al., 2015; Schleich et al., 2017) by providing evidence on the impact of high-frequency energy consumption information. More specifically, we complement the existing line of work by analyzing the impact of an In-Home Display using a field experiment that involves a larger sample size and a longer time period than is typical in this literature. This is important because larger sample sizes tend to yield more reliable estimates, and the general finding is that studies using larger samples document smaller effect sizes; see Delmas et al. (2013), Karlin et al. (2015, p. 1215), and also Agarwal et al. (2023). In our study, we monitor the daily amount of gas and electricity consumed by 802 households for a period of, on average, ten months post-treatment (and almost four months pre-treatment). In-Home Displays are expensive, and hence there is a trade-off between sample size and the credibility of the inference methods used. Earlier studies used either random assignment to treatment and control but with relative small sample sizes (typically less than 150 households; see for example Aydin et al., 2018; Westskog et al., 2015; Lynham et al., 2016; Schleich et al., 2017), or they had large sample sizes at the cost of using less rigorous designs. Examples of the latter include pilots relying on essentially before/after comparisons of the behavior of households that had purchased an IHD, or on comparing energy use between households that had or had not purchased an IHD; see for example Faruqui et al. (2010), McKerracher and Torriti (2013) and Ehrhardt-Martinez et al. (2010). The RCT by Schultz et al. (2015) is a notable exception as it monitors the energy consumption of about 431 households, but only for a relatively short period of time (up to three months after the date of installation of the IHD).

Second, while the bulk of the literature on feedback focuses on the impact on households' electricity consumption, we have access to households' day-to-day consumption of both electricity and natural gas. While electricity is the universal source of energy for running appliances, in many countries cooking, space heating and water heating are fueled by natural gas, propane, or oil. Many of the studies focusing on electricity consumption feedback are thus not able to estimate the impact on the latter behaviors, which typically account for the bulk of residential energy consumption. The scarcity of research on the impact of energy consumption feedback on *total* residential energy consumption (including cooking and heating) is evidenced by the fact that two of the four above-cited review articles focus on studies estimating the impact of just electricity (see Karlin et al., 2015; Khanna et al., 2021), while the other two, by Zangheri et al. (2019) and Agarwal et al. (2023), contain too few studies focusing on both gas and electricity to allow for a separate analysis.

Third, and related to the second, the overall effectiveness of providing feedback may depend on whether just one type of residential energy consumption is targeted, or all. On the one hand, the literature on moral licensing (see for example Tiefenbeck

[1] Smart meters are digital energy meters that track, real time, the amount of natural gas and electricity consumed at every moment, and allow access to this real-time data both in-home as well as remotely. Smart meters were introduced in the Netherlands in 2005, and by the time of the start of our RCT almost 40% of the Dutch households were equipped with such an intelligent energy metering system (RVO, 2018).

et al., 2013) suggests that providing feedback on just one type of energy (say electricity consumption) may invite reduced performance on another (e.g., more gas consumption); similar predictions can be derived from the literature on the mental costs of effort (for example because attention budgets are limited and/ or the costs of attention are convex; see Shenhav et al., 2017; Trachtman, 2024). On the other hand providing feedback on the use of just one activity or energy type may also give rise to positive spillovers on other environmental behaviors, for example if there are complementarities in energy consumption that result in improved savings in multiple domains. An example in point is Tiefenbeck et al. (2018) who estimated the impact of real-time feedback on resource usage for one specific energy-intensive activity, showering. Real-time feedback was provided on the amount of water and energy used as well as on water temperature. Tiefenbeck et al. (2018) find that providing feedback resulted in a 22% reduction in the amount of water and energy used per bath. But they also found, importantly, that the savings in electricity were five to six times larger than what was achieved, among the same population, using just aggregate electricity consumption feedback (as documented by Degen et al., 2013).[2] More generally, Zangheri et al. (2019) document that providing feedback on two energy sources (i.e., both electricity and natural gas) results in larger savings for each of the two than if just one source is targeted. Based on a meta-analysis they find that providing feedback on just electricity or just heating induces savings of, respectively, 8.3% and 4.6%, whereas the average reduction obtained with feedback on both electricity and heating equals 9.1%. Our experimental setup does not allow us to test how the IHD's effectiveness would have changed if it had provided feedback on just one of the energy types. Yet it is interesting to note that while we document sizable treatment effects for natural gas, the savings on electricity are considerable lower than the average savings of about 5% as documented in the extant literature — even though the results of Zangheri et al. (2019) suggest that the combined feedback should have resulted in a larger impact on electricity than typically documented by studies focusing on just electricity.

Fourth, and possibly associated with the previous, our study also takes a step towards assessing the effectiveness of providing appliance- or activity-specific feedback. Our study was implemented in the Netherlands, where natural gas is used for just three purposes: space heating, heating water and cooking. The activities associated with the latter two – showering and cooking – are easily observable within the household, and hence the energy intensity of each of these activities (and hence also the potential for saving energy) can straightforwardly be inferred from the changes in the amount of gas consumed as displayed by the IHD. This also means that any remaining gas consumption can reliably be attributed to space heating as well; see also CER (2011) and Harold et al. (2015). This is especially relevant because space heating, showering, and cooking are among the most energy-intensive uses — those whose energy intensities households tend to underestimate most, and hence those with the largest potential for energy savings; see Attari et al. (2010). As such, our paper also contributes to the insights provided by the emergent literature assessing the impact of appliance- and/or activity-specific feedback. As discussed above, Tiefenbeck et al. (2018) find strong evidence of the effectiveness of increasing the salience of the resource intensity of one specific activity (showering). These insights complement those of Gerster et al. (2021), who use a machine learning algorithm to provide targeted feedback, with a one-day delay, on not just one activity but on the usage of five of the most energy-intensive electric appliances in the household (the refrigerator, dish-washer, washing machine, dryer, and oven) as well as on the energy consumption of the remaining "always on" appliances. They find that increasing the salience of these six (groups of) appliances results in an additional 5 percentage points reduction in electricity consumption above and beyond what is achieved by just providing aggregate electricity feedback. In our setup households receive real-time (rather than delayed) energy consumption feedback that is easily attributable to showering, cooking and space heating, and hence our study provides a useful complement to the more targeted but less immediate feedback as provided by Gerster et al. (2021). And compared to Tiefenbeck et al. (2018) our IHD resulted in energy savings that are 1.5 times as high, possibly because it allowed participants in our study to straightforwardly update their beliefs about the energy intensity of showering but also of the other two gas-fueled activities — space heating and cooking.[3]

Fifth, our study also adds to the broader literature on the effectiveness of behavioral interventions in reducing energy consumption including social comparisons (Schultz et al., 2007; Allcott and Rogers, 2014; Pellerano et al., 2017; Brülisauer et al., 2020; Andor et al., 2020), changes in (the salience of) energy prices (with or without additional interventions like social comparisons, see Ito, 2014; Jessoe and Rapson, 2014; Pellerano et al., 2017), and changing the salience of the negative societal consequences of energy consumption (like environmental and health impacts; see Asensio and Delmas, 2015). Based on the meta-analyses of the effectiveness of various types of interventions (Faruqui et al. (2017) on financial incentives, Karlin et al. (2015) on energy consumption feedback, and Abrahamse et al. (2005) and Andor and Fels (2018) on social comparisons, goal setting, commitment devices and labeling), non-monetary interventions are effective but less so than financial ones; see also Khanna et al. (2021). The paper by Jessoe and Rapson (2014) is of particular interest as it shows how energy consumption feedback can increase the effectiveness of a pricing instrument: the electricity savings induced by peak-hour pricing are more than twice as high when accompanied with real-time feedback via an IHD. Our focus on the impact of real-time energy consumption feedback is especially relevant because of the increased availability of smart meters in developed countries, and then especially in the EU and the US. In 2021 the technology's residential penetration rates were 56% and 69% in these regions, respectively (Sovacool et al., 2021; Vitiello et al., 2022), and the numbers are expected to continue to increase if only because the EU's 2019 Energy Directive requires all member states to ultimately equip at least 80% of their households with a smart metering system (European Union, 2019).

[2] Interestingly, in a companion paper Tiefenbeck and co-authors found similarly large effects of providing water consumption feedback among hotel guests — that is, in a situation where users do not have any private financial incentives to save water; see Tiefenbeck et al. (2019).

[3] Tiefenbeck et al. (2018) find that providing feedback on showering results in 1.2 kWh energy savings per day, while we find that the IHD caused the daily amount of gas consumed to decrease by 0.185 m^3; see Table 4 in Section 3. Using the conversion rate of 1 m^3 of natural gas being equal 9.769 kW, this translates into 1.81 kWh of gas saved per day – 1.5 times the savings reported by Tiefenbeck et al. (2018).

Fig. 1. The GEO TRIO II In-Home Display.
Note: The In-Home Display used in this study. Key features are the information about the consumption of electricity and natural gas, presented simultaneously, either at this instant (in the tab labeled "Nu") or cumulatively over the day (if the tab labeled "Vandaag" is activated). The gauge and the size of the flame reflect the relative intensity of usage of respectively, electricity and natural gas. The LED light on the top of the IHD also reflects the intensity of current usage.

The set-up of this paper is as follows. Section 2 describes the intervention as well as the actual implementation of the RCT, including the identification strategy. Section 3 presents the estimated impact of the IHD on the consumption of both natural gas and electricity, and Section 4 presents the results of a survey aimed at uncovering the mechanisms via which the observed savings have been realized. Section 5 concludes.

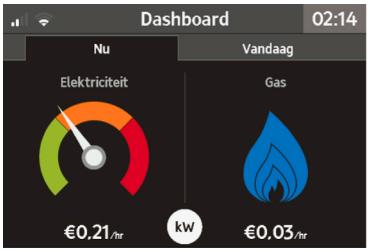
2. Experimental design

2.1. Background and treatment description

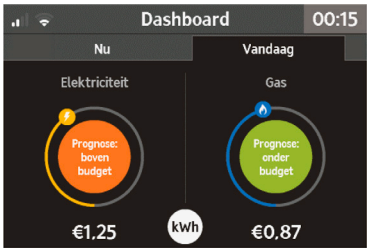
Households typically consume energy without knowing how much they actually use at every instant, nor what the financial consequences are (Abrahamse et al., 2007; Attari et al., 2010). This lack of information may cause households to consume too much energy — directly, in the form of their actual consumption decisions, but also indirectly, in the form of a (sub-optimal) choice of the types of technologies and appliances they purchase and the energy conservation measures they forego (Allcott et al., 2014; Carlsson et al., 2021). Energy bills are sent out regularly but at relatively low frequency, and they typically only provide information about the total amount of electricity and/or gas consumed during the billing period (Jessoe and Rapson, 2014). This makes it difficult for households to derive information about the financial consequences of different household activities and/ or the use of specific appliances. In the Netherlands, for instance, energy is billed just once a year (with households paying monthly installments of one-twelfth of the estimated amount they have to pay over the entire billing period). Smart energy meters offer households the possibility to receive high-frequency information on their energy consumption via, among other devices, in-home displays (IHDs). These IHDs, also referred to as energy monitors, do not only allow households to obtain better insight into their energy consumption patterns, but also to access and digest that information whenever it suits them.

In this study, we experimentally estimate the impact of the real-time energy consumption feedback provided by one such IHD, the GEO TRIO II (see Fig. 1), on the amount of natural gas and electricity consumed by (a subsample of) Dutch households. The usage data are visualized by means of a simple color-coded dashboard; see Fig. 2. The device can display current energy consumption (for natural gas and electricity separately), but it also has a self-learning algorithm to predict the amount of gas and electricity that will be consumed during (the remainder of) the day. The various types of information can be displayed in either physical units or monetary costs, as well as with or without comparison to a "daily target budget" set by the user. In addition, the multi-colored LED light fitted to the top of the display signals the level of current electricity usage (green, orange and red, depending on the household's current energy usage relative to its "normal" energy consumption, as determined by the algorithm).

The IHD can affect households' energy consumption via a number of mechanisms. First, if placed in a clearly visible spot (such as in the kitchen or living room), the IHD can raise the salience of the amount of energy consumed (Schwartz et al., 2013; Jessoe and Rapson, 2014), and this effect may be reinforced by the LED light on top of the display. And the information on the display itself can increase the salience of the *costs* of energy consumption. Brounen et al. (2013) found that only about half of Dutch households were able to provide a by and large accurate estimate of their monthly energy payments; the IHD may help improve households' knowledge of how costly it is to use gas and/ or electricity (Sexton, 2015). Second, the IHD may help households to gain insight into the relative energy intensity of the various types of energy use. Households typically underestimate the amount of energy consumed by the most energy-intensive activities (Attari et al., 2010). Having access to real-time information on energy consumption may enable them to update their knowledge of how much energy is associated with the use of specific appliances (if only by switching them on and off, and seeing how this affects the readings on the IHD), and hence how costly it is to run them. This type of learning may be especially relevant for the consumption of natural gas, as gas-fueled activities are typically more energy-intensive than most of the electricity-powered activities. Even though our IHD only provides information about the aggregate amount consumed,



(a) The current (="Nu") energy usage tab of the IHD's home screen, here shown displaying the amount of money currently being spent on electricity and gas.



(b) The today's (="Vandaag") energy consumption tab of the IHD's home screen, here shown displaying the amount of money spent on energy today, and accompanied by a prognosis of whether today's consumption will stay within the household's targeted daily budget.

Fig. 2. Key functions of the energy in-home display.

real-time, of each energy type, establishing the energy-intensity of the various uses of gas is fairly straightforward, because natural gas is used for just three activities that are very visible (or easily detectable) within the household (Harold et al., 2015): space heating, showering, and cooking. Establishing the link between each of the activities and the amount of natural gas consed is even more straightforward if the IHD is centrally located in the home — think of the kitchen or living room. And this learning may also increase the household's self-efficacy, if the IHD increased the household's confidence in their ability to reduce their energy consumption (Bandura, 1977; Lauren et al., 2016).

Third, the IHD offers the possibility of setting daily consumption target levels (or daily budgets) for gas and electricity consumption, and hence may act as a soft commitment device for saving energy (Bandura, 1988; Benhabib and Bisin, 2005; Andor and Fels, 2018). Finally, the presence of the IHD may also act as a constant reminder of the (societal) relevance of saving energy, thus affecting the household's attitudes towards taking energy-saving actions (Abrahamse et al., 2007; Hargreaves et al., 2010). We estimate the overall impact of the IHD on gas and electricity consumption using our participant households' smart meter data, but we also probe the relative importance of the above mechanisms by means of a survey.

2.2. Recruitment process and data description

To analyze the impact of the IHD on residential energy consumption, we partnered with seven organizations in the Netherlands to recruit a large pool of households interested in receiving an IHD, free of charge, for a period of approximately 12 months. Our partners included social housing- and tenant associations, municipalities, and agencies aiming to foster energy savings. Here, we define a region (also referred to as location) as the cluster of municipalities (as identified by two-digit zip codes) serviced by an implementer organization. The recruitment campaign took place between September 2017 and June 2019, and consisted of leaflets distributed in public places, advertisements in local newspapers, and social media messages. The communication was focused on informing the reader that (i) displays were available that provide insight into actual energy usage which can help households save energy and reduce their energy bill, and (ii) these displays were to be allocated, by means of a lottery, among all those interested in participating in the lottery, if a number of restrictions was met. Participation was restricted to households that lived in a gas-heated dwelling in a region serviced by our partner organizations, and that did not have the intention to relocate to a new home within the next year. In addition, the dwelling had to be equipped with a smart meter that could be accessed remotely by their energy company, and participant households had to be willing to agree to sharing their energy consumption data for research purposes for a period of one year independent of whether they won an IHD or not.

The participants in this study were thus households interested in receiving an IHD; which of these households were offered an IHD was determined by means of a lottery. Our identification strategy is thus a random rejection (or oversubscription) design; see Angrist et al. (2002). The lottery was location- as well as recruitment-period specific. Using the term 'batch' to denote the stock of households in a location that had applied within a specific time period, we waited until the number of applicant households in a batch was large enough before randomizing the households with IHD-compatible smart meters in that batch into the group that would be offered an IHD (the treatment households) or the group that would be informed that they were not to receive an IHD (the control group). In case of perfect treatment compliance, the Average Treatment Estimate (ATE) for this oversubscription design is effectively a Local Average Treatment Effect (LATE) estimate (see Imbens and Angrist, 1994) as it provides an estimate of the treatment impact on a subpopulation of Dutch households — those that were interested in receiving the IHD. IHDs are expensive (about €100 a piece at the time), and hence the choice for an oversubscription design is warranted because of two reasons. One, randomly allocating IHDs to a sample of households interested in receiving an IHD maximizes the statistical power to estimate the impact of the IHD among those in the population who are interested in saving energy via an IHD. Two, the impact on those interested in receiving the IHD is more policy relevant than the average effectiveness among the general population in the Netherlands, as the households in our sample are more likely to be representative of the ones willing to purchase a (subsidized) IHD in case of policy scale-up; cf. List (2024). Having said that, an oversubscription design brings challenges of its own, and also the actual implementation was

not straightforward in all respects. Before presenting our identification strategy, we discuss five issues that complicated our impact evaluation.

First, compliant with the EU's GDPR we asked our participants for informed consent to allow us to download their natural gas and electricity consumption data, and hence the participant households were aware of the fact that their data were to be analyzed for research purposes. Our study is thus not a natural experiment as defined by Harrison and List (2004), and the knowledge that one's behavior is monitored may affect behavior even absent any actual treatment — the so-called Hawthorne effect (Bloombaum, 1983; Levitt and List, 2007). Informing households of the purpose of the project was necessary because we needed access to applicant households' smart meter data independent of whether they were to receive an IHD or not. While this setup has the advantage that our study does not suffer from attrition, we cannot rule out that our results may have been affected by the Hawthorne effect. Having said that, we think the effect is limited, because of the following reasoning.

If present, the Hawthorne effect is likely to have affected the behavior of both experimental groups and hence does not, by itself, bias the treatment estimate. However, possible Hawthorne effects are likely to have been stronger for the treatment households, as the physical presence of the IHD may have acted as a reminder of the fact that their energy consumption was monitored by researchers (cf. Hargreaves et al., 2010; Schwartz et al., 2013); our impact estimate would then be an overestimate of the actual impact of having an IHD. While the Hawthorne effect (if present) is thus likely to have been smaller for the control group, other mechanisms may have been at play too that can have affected the amount of energy consumed — and then especially so for the control group. More specifically, having been informed that one is not to receive an IHD may have induced households in the control group to change their energy consumption behavior directly and/or indirectly. Control group households may have been disappointed when learning they were not going to receive an IHD, and the direct impact of this is, ex ante, ambiguous. On the one hand, this may have decreased the control group's intrinsic motivation to save energy. On the other hand, it may also have induced the control group households to save more energy — if the message that they had not won an IHD invited reactance: the so-called John Henry effect (see Saretsky, 1972). Whether the one response dominated the other, is an empirical question that we are, unfortunately, not able to answer. But then having been denied an IHD may also have had an indirect effect on the control group's energy consumption behavior — if the rejection induced (some of the) control households to purchase an energy monitor themselves. While the direct effect of the disappointment on control group behavior is ex ante ambiguous, the indirect effect is in the direction of our design resulting in an underestimate of the true effectiveness of receiving an IHD.

The direct and indirect effects do not just differ in the ex-ante ambiguity of their impact; they are also likely to differ in the extent to which they are long-lasting. The direct effects may have been sizable in the short run, but are less likely to have been so too in the longer run. Tiefenbeck (2016) finds that in field experiments on energy conservation, any Hawthorne effects fade quickly over time. Darby (2006, p.7) comes to a similar conclusion, as she writes that "[the Hawthorne effect is] typically less large with field studies carried out on a longer timescale and with larger samples". We have a baseline period with a median length of 94 days and an intervention period with a median length of 264 days, and hence we are confident that any research-induced change in behavior – the overall Hawthorne effect, but also the (potentially) stronger Hawthorne effect for the treatment group and the disappointment effect among the control group – are unlikely to have substantially affected our treatment estimates. If anything, our random rejection design may have resulted in an underestimate of the actual impact of the IHD because of the indirect effect; that is, if some of the households in the control group decided to acquire an energy monitor themselves.

Second, while straightforward on beforehand, implementing our design was more complicated than anticipated as not all smart meter types turned out to be compatible with our IHD. Table 1 presents information, for each of the 25 batches, on the number of applicant households in each batch, how many of them had smart meters that were compatible with our IHD, and how many where allocated to treatment and control. Regarding the IHD-compatibility of households' smart meters, a sizable share of households were equipped with meter types that were IHD-compatible, which could thus be randomized into treatment and control. Despite the manufacturer's claim that the IHD could be paired with any type of smart meter in the Netherlands, a non-negligible share of applicant households turned out to have smart meters of one of two types that were not compatible with our IHD: the DSMR 2.0, and ISKRA. This complicated the design of our study because of two additional issues: (i) the recruitment of participant households was quite slow at the start (and also slower than anticipated; see the timing and number of households in each of the batches presented in Table 1), and (ii) our implementing partners faced time and budgetary pressures to finalize the roll-out of the IHDs.

Regarding the two types of IHD-incompatible smart meters, we decided to exclude all households with DSMR 2.0 m, as this meter type was introduced in the Netherlands prior to the nationwide roll-out of smart meters (in 2015). Before the start of the nationwide smart-meter roll-out households could request to receive a smart meter as part of a pilot program, and hence households that were most eager to acquire a smart meter were thus more likely to end up with a DSMR 2.0. We also could afford dropping these households, as only few of the applicant households had this type of smart meter. Whether households with ISKRA meters were unlikely to provide an unbiased counterfactual for the amount consumed by the treatment households was less clear-cut. The ISKRA meters only became available after the nationwide roll-out of the smart meter program had started, when there were already quite a number of different meters on the market. During the nationwide roll-out all smart meters were offered for free, but households were not allowed to choose which meter type they wished to receive. We contacted the organization that oversaw the smart-meter roll-out for details. They stated that network operators typically had multiple types of smart meters on stock to avoid becoming overly dependent on a single supplier, and hence that the allocation of IHD-compatible and IHD-incompatible smart meters was quasi-random: the type of meter a household ended up with was determined by the type of meters the network operators had on stock at the time. The ISKRA meter turned out to be quite prevalent in our sample: in total, 241 of the 810 households (including outliers; see below) in our sample had an ISKRA meter.

Table 1
The number of participant households by date and region of recruitment, and by smart meter type and treatment status.

Date randomization	Location							
	1	2	3	4	5	6	7	Total
Jan 11, 2018								
Total N				19				19
N Non-compatible, control				0				0
N Compatible, control				9				9
N Compatible, treatment				10				10
Apr 23, 2018								
Total N						48		48
N Non-compatible, control						0		0
N Compatible, control						25		25
N Compatible, treatment						23		23
June 11, 2018								
Total N				8			69	77
N Non-compatible, control				0			20	20
N Compatible, control				3			15	18
N Compatible, treatment				5			34	39
June 23, 2018								
Total N	20		14			7[1]		41[1]
N Non-compatible, control	2		10			0		12
N Compatible, control	8		0			4		12
N Compatible, treatment	10		4			3[1]		17[1]
July 6, 2018								
Total N				10		9		19
N Non-compatible, control				0		1		1
N Compatible, control				5		4		9
N Compatible, treatment				5		4		9
Sep 24, 2018								
Total N				12			7	19
N Non-compatible, control				0			2	2
N Compatible, control				7			0	7
N Compatible, treatment				5			5	10
Oct 24, 2018								
Total N	4					25[1]		29[1]
N Non-compatible, control	0					7[1]		7[1]
N Compatible, control	2					4		6
N Compatible, treatment	2					14		16
Nov 30, 2018								
Total N						63[1]	9	72[1]
N Non-compatible, control						29	2	31
N Compatible, control						2	0	2
N Compatible, treatment						32[1]	7	39[1]
Feb 15, 2019								
Total N				9			12	21
N Non-compatible, control				0			3	3
N Compatible, control				5			0	5
N Compatible, treatment				4			9	13
July 10, 2019								
Total N		62		8	185[1]	137[4]	41	433[5]
N Non-compatible, control		4		2	0	45[2]	17	68[2]
N Compatible, control		22		2	88	16[1]	0	128[1]
N Compatible, treatment		36		4	97[1]	76[1]	24	237[2]
04 Aug 2019								
Total N		12	11		1			24
N Non-compatible, control		1	6		0			7
N Compatible, control		6	5		1			12
N Compatible, treatment		5	0		0			5
Total								
Total N	24	74	25	66	186[1]	289[7]	138	802[8]
N Non-compatible, control	2	5	16	2	0	82[3]	44	151[3]
N Compatible, control	10	28	5	31	89	55[1]	15	233[1]
N Compatible, treatment	12	41	4	33	97[1]	152[3]	79	418[4]

Note: The number of participant households with IHD-compatible smart meters and those with incompatible unupdatable (ISKRA) meters, per location and per date of allocation to treatment and control. Note that we were able to install an update among all 87 of the ISKRA-metered households in Location 5 with batch date July 10, 2019, so that their meters became IHD-compatible. We have, in total, 810 applicant households, eight of which we consider to be outliers because their estimated yearly amount of electricity purchased was above 9000 KWh, which is well above the 8150 kWh consumed by the 95th percentile of the group of top residential energy users in the Netherlands (those households consisting of five or more household members living in large detached houses in the Netherlands). The figures in normal font in each cell denote the number of households to be included in our main analysis; the number of outliers in a cell is indicated by the associated superscript. E.g., the cell with 7^1 (in Location 6, with randomization date June 23, 2018) indicates that this batch consisted of 8 households, one if which meets the definition of an outlier provided above. From the remainder of the information for this batch, the reader can infer that this outlier had originally been assigned to the treatment group.

We proceeded as follows. We requested the manufacturer of the IHD to develop an update for the ISKRA meters to make them IHD-compatible. Luckily the manufacturer succeeded to produce an update just in time to install the update among all 87 ISKRA-metered households in the July 10, 2019 batch of Location 5, so that they could be randomized into treatment and control; see Table 1. Eventually that left us with 154 unupdated ISKRA meters in our sample. But because of the timing of the recruitment process, we had to decide whether the ISKRA-metered households could provide a proper counterfactual for the amount of energy consumed by the treatment households (i.e., whether they could be included in control group) much sooner — early on in the recruitment process, when it was still very much uncertain whether an update could be developed. As stated above, according to the oversight organization the process of the roll-out of smart meters thus had been such that it was quasi-random which households ended up with an ISKRA meter. We tried to formally test this using two types of analyses — whether households with and without IHD-compatible smart meters had similar house- and household characteristics (including the amount of gas and electricity consumed during the baseline period), and also whether there were no systematic differences in the amounts of energy consumed by the two types of households; see Appendix A.1. As shown in Table A1 in that appendix, households with unupdatable ISKRA meters were very similar in terms of their observable characteristics to those with IHD-compatible meters. Only one of the twenty house- and household characteristics shows up significantly (ISKRA meters were more prevalent in terraced houses; $p = 0.073$), but this difference does not survive multiple hypotheses testing (MHT); see Table B2 in Appendix B (Supplementary Materials).[4] A1 also shows that none of the normalized differences in the characteristics of the two types households exceed the critical threshold of 0.25 standard deviations that is typically used in evaluating between-group similarity (Imbens and Wooldridge, 2009; Imbens and Rubin, 2015). We thus cannot reject the null of the two types of households being similar in terms of their house- and household characteristics, and hence it is plausible that the parallel trend assumption is likely to be met too. We explicitly tested for this in Fig. A.1 in Appendix A.1. This figure plots the difference in the amounts of gas and electricity consumed between ISKRA-metered households and those with IHD-compatible smart meters in the four months prior to and in the first four months after the start of the intervention.[5] None of the differences are significant at the 10% or lower.

We thus find no evidence to reject the claim of the allocation of smart meters having been quasi-random. Still, we would have preferred to exclude all ISKRA-metered households from the sample and continue recruiting just households with IHD-compatible smart meters. This turned out to be unacceptable to our implementing partners because of the two reasons mentioned above: the cumbersome and hence costly recruitment process (especially in the first six months since the start of the project; see Table 1), and the political and budgetary pressure on our implementing partners to finalize the roll-out. In view of the evidence regarding the quasi-random allocation of smart meter types we decided to keep all ISKRA-metered households in the control group, and randomly allocate the IHD-compatible households to treatment and control using a higher (lower) probability of these households ending up in treatment (control); see Table 1 and also Section 2.3 for more information. By so doing, we maximize statistical power when using the full sample of households (as including the ISKRA-metered households results in half of the sample ending up in treatment and half in control), while we can also test whether the estimates obtained are similar when restricting the sample to just those households with IHD-compatible smart meters — but then with a 2:1 ratio of households in treatment and control (rather than 50-50). We present the outcomes of this alternative analysis as one of the robustness tests; see Section 3.2 as well as Appendix A.2.2.

Third, we had, in total, 810 applicant households for which we were able to retrieve their energy consumption. Of these, we excluded eight households from our main analysis as their estimated yearly amount of electricity consumed was above 9000 KWh, which is well above the 8150 kWh that is consumed by the 95th percentile of the subpopulation of households in the Netherlands with the highest electricity consumption — those with five or more household members living in a large detached house.[6] The number of outliers in a batch is indicated in Table 1 by means of superscripts. For example, "7^1" in the 23 June 2018 batch in Location 6 indicates that batch consisted of seven households that are included in our main analysis, and of one household that is considered to be an outlier according to the above definition; from the number "3^1" elsewhere in that batch one can infer that this outlier was a household with an IHD-compatible smart meter and that it had been allocated to the treatment group. As can be inferred from the summary statistics at the bottom of Table 1 the eight outliers were equally distributed between treatment and control. Our experimental sample thus consists of 802 households which were recruited between January 11, 2018 and August 4, 2019, and 418 of which were offered an IHD. Simply omitting outliers may be considered too drastic, and hence we will test whether our results are robust to using alternative methods of dealing with outliers too; see Section 3.2 and also Appendix A.2.4.

Fourth, we aim to measure the impact of the IHD on the daily amount of gas and electricity *consumed* by each household, but smart meters register only the daily (net) amount of gas and electricity *purchased*. For gas, the amount purchased equals the amount consumed, and the same holds for electricity for those households without solar panels. For households with solar panels, however, the amount of electricity consumed is equal to the amount generated by the household's solar panels plus the net amount of electricity the household purchased from the grid; smart meters record the latter, but not the former. As the share of households owning solar panels at baseline is strongly balanced (with $p = 0.385$ according to a standard t-test and with MHT-corrected p-values

[4] We implemented four different MHT corrections to test the significance of the differences for each of the twenty characteristics: Anderson sharpened q (Anderson, 2008), Bonferroni–Holm (as implemented in Jones et al., 2019), Sidak–Holm (Abdi, 2007), and Westfall–Young (Westfall and Young, 1993). None of the twenty differences has an MHT-adjusted p-value below 0.781 across all four different tests; for details see Table B2 in Appendix B (Supplementary Materials).

[5] For the four months prior to the start of the intervention the comparison is between amounts of natural gas and electricity consumed by the ISKRA-metered and by all IHD-compatible households; in the four months after the start of the intervention the IHD-compatible households that were assigned to the treatment group were dropped from this parallel-trend analysis.

[6] For details see https://www.cbs.nl/nl-nl/cijfers/detail/83882ned; website consulted on 10-7-2020.

of 0.975 or higher; see Table 2, and also Table B3 in Appendix B, Supplementary Materials), just having data on the net amount of energy purchased for the subgroup of households with solar panels is unlikely to bias our impact estimates, but it does lead to less precise estimates (Pischke, 2007).

To increase precision we estimated the amount of electricity produced on each day by each of the households with solar panels as follows. A household's electricity production on a particular day depends on the amount of solar energy inflow on that day in the household's location, as well as on the electricity production capacity of the household's solar panel system. We estimated the household's production capacity by (i) measuring the maximum amount of electricity the household supplied to the grid in any hour on any day in our data set, and (ii) estimating the amount of electricity consumed by each household's "always on" appliances by calculating the average amount of electricity that household consumed, per hour, between 3 and 5 a.m., over the entire period for which we have data. Assuming that the amount of electricity used by the "always on" appliances is by and large constant throughout the day (cf. Gerster et al., 2021), a household's solar panel electricity production capacity is then equal to the sum of (i) and (ii), and our estimate of the amount of electricity generated at any moment by each household is then equal to the household's solar system's production capacity multiplied by the location's solar energy inflow at that moment divided by the solar inflow that was registered when the peak capacity was measured. We find that households with a solar panel installation consume, on average, 31 percent of the amount of electricity they produce. This is in line with the estimates by Luthander et al. (2015) who find that the share of own solar electricity production consumed lies between 17 and 38 percent among German households. We analyze the relevance of correcting for the amount of electricity produced using solar panels as one of the robustness tests; see Section 3.2 and also Appendix A.2.6.

Fifth, the households' smart meters register the daily (net) amount of gas and electricity purchased by each household, and we downloaded this data for a period of, on average, 14 months. The start of the intervention period for a household that was offered and accepted the IHD is defined by the moment at which the IHD was installed in their home. Unfortunately, IHDs were not installed in 74 of the 418 households (or 17.7%) that had been assigned to the treatment group. Failure to install the IHD was caused by technical difficulties, but also because of households having changed their mind and decided to decline the offer. We use a batch's median installation date as the starting date of the intervention period for the treatment households that did not end up with an IHD. The median time difference between the date we started receiving the energy consumption data and the time of installation of the IHD was 94 days, and the median length of the intervention period was 264 days. Our panel data set covers the period between December 7, 2017, and June 14, 2020, but it is unbalanced because of our rolling recruitment design as summarized in Table 1. In addition, errors in the data collection resulted in us not having access to all energy consumption data of each individual household. In total, our panel data set consists of more than 340,000 observations for the amount of gas consumed per day, and about the same number for electricity.

2.3. Randomization process and balance tests

As can be inferred from Table 1, sign-up was relatively slow especially at the beginning of the campaign. To avoid letting applicant households wait too long to learn whether or not they were to receive an IHD, the randomization was implemented per location and per subperiod, where, depending on the share of unupdatable ISKRA meters in the batch, we overassigned IHD-compatible households to the treatment group to achieve a 50-50 allocation to treatment and control. Randomization per batch is important because we did not want to keep households waiting too long, but also because it may be non-random which households in a region signed up early or late; the random assignment of the households with IHD-compatible smart meters to treatment and control was thus stratified both on timing and location. Using $\lceil x \rceil$ to denote the largest integer equal to or below x, the probability π_{ij} of IHD-compatible household i in batch j to be allocated to the treatment group was set equal to $\min\left[\frac{\lceil 0.5N_j \rceil}{N_j - UI_j}, 1\right]$, where N_j denotes the total number of applicant households in randomization batch j and UI_j the number of applicant households with an unupdatable ISKRA meter in that batch. As can be inferred from Table 1, $\lceil 0.5N_j \rceil$ was larger than $N_j - UI_j$ for just one batch — that of June 23, 2018, in Location 3.

Table 2 presents the descriptive statistics of each of the two experimental groups (see columns (1) and (2)) as well as the outcomes of the associated balance tests (in columns (4) and (5)); column (3) of Table 2 presents the characteristics of the overall sample of households that participated in our study. Most of the participant households consisted of either one or two persons (30% and 57%, respectively), about a quarter had at least one household member aged over 65, and 34% had at least one child below 18. In only very few households primary education was the highest level of educational attainment among all household members, and about half of the households lived in either apartments or terraced houses. About 44% of the participant households were owner-occupants, about 38% lived in homes that were built after 1974, and about 41% were owners of a solar power system at baseline. Finally, households used on average about 1.66 m^3 of gas and 7.72 kWh of electricity per day; the average value of energy consumption was €2.62 per day. As shown in Table B1 in Appendix B (Supplementary Materials), the households in our sample are about equally educated as the average person in the Netherlands, and are equally likely to have children as the average household in the Netherlands. Regarding housing, household in our sample are slightly more likely to live in a home that is rented, smaller and older than that of the average household in the Netherlands. Most importantly (but not surprisingly), households in our sample are much more likely to have a solar panel system on their rooftop than the average household in the Netherlands.

We conducted balance tests between the treatment and control groups to determine whether our randomization procedure resulted in two groups that were similar on observables. Columns (1) and (2) of Table 2 present the sample means of the control and treatment group, and columns (4) and (5) display the absolute and normalized differences between the two means, respectively. As

Table 2
Descriptive statistics and balance tests.

Variable	(1)		(2)		(3)		(4)	(5)
	Control group		Treatment group		Total		Difference	Normalized diff.
	N	Mean/SE	N	Mean/SE	N	Mean/SE	(1)–(2)	(1)–(2)
1 person hh	384	0.318	417	0.288	801	0.302	0.030	0.065
		(0.024)		(0.022)		(0.016)		
2 person hh	384	0.552	417	0.580	801	0.567	−0.028	−0.057
		(0.025)		(0.024)		(0.018)		
> 2 hh members	384	0.130	417	0.132	801	0.131	−0.002	−0.005
		(0.017)		(0.017)		(0.012)		
Elderly hh	384	0.229	417	0.293	801	0.262	−0.063**	−0.144
		(0.021)		(0.022)		(0.016)		
Household with children below 18	384	0.359	418	0.323	802	0.340	0.036	0.077
		(0.025)		(0.023)		(0.017)		
Just primary education	384	0.026	417	0.026	801	0.026	−0.000	−0.002
		(0.008)		(0.008)		(0.006)		
Secondary or lower tertiary education	384	0.534	417	0.525	801	0.529	0.009	0.017
		(0.025)		(0.024)		(0.018)		
Higher tertiary education	384	0.440	417	0.448	801	0.444	−0.008	−0.017
		(0.025)		(0.024)		(0.018)		
Apartment	384	0.227	417	0.266	801	0.247	−0.040	−0.092
		(0.021)		(0.022)		(0.015)		
Terraced house	384	0.284	417	0.242	801	0.262	0.042	0.095
		(0.023)		(0.021)		(0.016)		
Corner house	384	0.195	417	0.185	801	0.190	0.011	0.027
		(0.020)		(0.019)		(0.014)		
Semi-detached house	384	0.169	417	0.163	801	0.166	0.006	0.017
		(0.019)		(0.018)		(0.013)		
Detached house	384	0.112	417	0.125	801	0.119	−0.013	−0.039
		(0.016)		(0.016)		(0.011)		
Other (farmhouse, houseboat)	384	0.013	417	0.019	801	0.016	−0.006	−0.049
		(0.006)		(0.007)		(0.004)		
Dwelling owned by the household	384	0.435	417	0.446	801	0.441	−0.011	−0.022
		(0.025)		(0.024)		(0.018)		
House built after 1974	383	0.366	417	0.384	800	0.375	−0.018	−0.037
		(0.025)		(0.024)		(0.017)		
Owned photovoltaic cells at time of randomization	384	0.393	418	0.423	802	0.409	−0.030	−0.061
		(0.025)		(0.024)		(0.017)		
Baseline average daily gas consumption (m^3)	384	1.610	418	1.702	802	1.658	−0.092	−0.045
		(0.101)		(0.105)		(0.073)		
Baseline average daily electricity consumption (kWh)	384	7.668	418	7.775	802	7.724	−0.107	−0.024
		(0.219)		(0.223)		(0.156)		
Baseline average daily energy expenditures, in Euros	384	2.580	418	2.661	802	2.622	−0.081	−0.044
		(0.090)		(0.094)		(0.065)		

Notes: Overall sample statistics (means and standard errors, in column (3)) and balance tests using absolute and normalized differences (in columns (4) and (5), respectively) for the mean characteristics of the treatment and control group households (presented in columns (2) and (1), respectively). The table was constructed using the iebaltab command of the STATA's ietoolkit package (DIME Analytics, 2019). *, **, and *** indicate that the standard t test of the difference presented in column (4) is significant at, respectively, the 10, 5, and 1 percent critical level. When correcting for Multiple Hypotheses Testing (MHT) all absolute differences are insignificant (with $p = 0.545$ or higher) according to a series of MHT tests (Anderson sharpened q, Bonferroni–Holm, Sidak–Holm and Westfall–Young; see Table B3 in Appendix B, Supplementary Materials). As shown in column (5), the absolute values of all normalized differences are all smaller than 0.25 standard deviations, the threshold above which imbalance is considered to be a concern (Imbens and Wooldridge, 2009; Imbens and Rubin, 2015).

is clear from column (4), the absolute differences between the two experimental groups are small, and only one difference is found to be statistically significant according to a standard t-test: the share of households with one or more elderly members is larger in the treatment group than in the control group ($p = 0.042$). This is, however, not likely to be a true positive as this difference does not survive correcting for Multiple Hypotheses Testing; see Table B3 in Appendix B (Supplementary Materials). Using four different types of corrections (Anderson sharpened q, Bonferroni–Holm, Sidak–Holm, and Westfall–Young), MHT-adjusted p-values for the difference in the share of elderly households range between 0.572 and 1.000, and also all differences in the other 19 characteristics have MHT-adjusted p-values that are 0.866, or higher. And as shown in column (5) of Table 2 none of the normalized differences are larger than the critical threshold of 0.25 that is typically used in this literature (Imbens and Wooldridge, 2009; Abadie and Imbens, 2011; Imbens and Rubin, 2015).[7]

2.4. Identification strategy

As stated in Section 2.2, with perfect compliance the difference in average consumption between treatment and control in an oversubscription design is, in fact, the treatment-on-the-treated estimate — the impact on the energy consumption of those households interested in receiving an IHD for a year (Angrist et al., 2002). Unfortunately, IHDs were not installed in 74 of the 418 households that had been assigned to the treatment group; see Section 2.2. That means that we can estimate two impacts for the IHD: the intention-to-treat (ITT) estimate as well as the Local Average Treatment Effect (LATE). In our oversubscription design with imperfect compliance the ITT estimate is the Average Treatment Effect (ATE) of the impact of the IHD on the amount of

[7] We also verified whether the treatment and control groups were balanced within each region. Of the (7 × 20 =) 140 bicomparison tests, only seven differences were significant at the 10% level, or lower – a number that is below the number of false positives one would expect given the 10% significance threshold.

energy consumed by households interested in receiving an IHD, and the LATE is the Average Treatment on the Treated (ATT), or the impact on the amount of energy consumed by those households that expressed interest in receiving the IHD and also ended up having one installed in their home (cf. Imbens and Angrist, 1994). The specifications used to estimate the ATEs and ATTs are presented, respectively, in Eqs. (1) and (2)–(3).

The regression model to estimate the Average Treatment Effect (ATE) is specified as follows.

$$y_{it} = \beta_{ATE} T_{it} + \delta X_{rt} + \alpha_i + \gamma_{rt} + \epsilon_{it}. \tag{1}$$

Here, y_{it} denotes the amount of energy consumed (gas, electricity, or the total amount of energy consumed) by household i on day t. T_{it} is a variable that is equal to one for all households in the treatment group from either the day onwards at which the IHD had been installed, or from the batch's median installation date (for those households in the treatment group that ended up not accepting the IHD), and zero otherwise. We include household fixed effects, α_i, as well as time fixed effects, γ_{rt}, where subscript r denotes the location household i lives in. Regarding the latter, the time fixed effects can be calendar day fixed effects that are sample-wide (so that they are the same for all seven locations), or they can be location-calendar day specific – hence the subscript r – to allow day-to-day shocks to differ between locations. X_{rt} is the region-specific vector of time-varying variables that may affect the demand for energy: daily average temperature (both in levels as well as squared, to allow for non-linearity) and the region-specific amount of solar radiation (cf. Asensio and Delmas, 2016). The values for these variables are obtained by identifying, for each household i, the nearest weather station of the Dutch Meteorological Institute. Next, ϵ_{it} denotes the error term, which is clustered either just at the household level, or also at the 2-digit zip code level for all households living in a region with at least one unupdatable ISKRA meter. Following Abadie et al. (2023), clustering the error terms for households within locations where unupdatable ISKRA meters were present takes into account the fact that not all households could be randomized into treatment and control by the experimenter, as we are relying in part on the quasi-random allocation of smart meter types across households (but see Appendix A.2.2 for the analysis using just IHD-compatible households). Because of the geographic component of the smart-meter roll-out, clustering at the municipality (i.e, two-digit zipcode) level is likely to be preferred over clustering at just the household level. Finally, β_{ATE} is the coefficient of interest — the Average Treatment Effect of offering the IHD to our sample of participant households.

Next, the Average Treatment on the Treated (ATT) is estimated using the following 2SLS model:

$$I_{it} = \rho T_{it} + \eta X_{rt} + \theta_i + \psi_{rt} + \mu_{it}, \text{ and} \tag{2}$$

$$y_{it} = \beta_{ATT} \hat{I}_{it} + \delta X_{rt} + \kappa_i + \lambda_{rt} + \nu_{it}. \tag{3}$$

In Eq. (2), successful installation of the IHD is captured by I_{it}, a dummy variable that is equal to 1 from the day onwards on which the IHD had been installed in household i, and zero otherwise. In essence this equation estimates the probability of the IHD ending up having been installed in a household, with that household's treatment status as instrumental variable; the other variables in Eq. (2) are the vector of weather controls (X_{it}) and the household (θ_i) and time (ψ_{rt}) fixed effects, where the time fixed effects may again either be calendar day or region-calendar day fixed effects. In Eq. (3) the variable of interest (y_{it}) is regressed on the estimated probability of the IHD having been installed in household i at calendar day t ($\hat{I}_{it}$, as predicted by Eq. (2)), on the vector of weather variables (X_{it}), and also on the household and time fixed effects (α_i and γ_{rt}). The coefficient of interest is β_{ATT}, which is the ATT (or LATE): the impact estimate of the treatment on those households in whose homes the IHD was successfully installed. Finally, μ_{it} and ν_{it} are error terms, clustered again at either the household level or also at the 2-digit zip code level for all households living in a region with at least one unupdatable ISKRA meter. Note that our instrument (T_{it}) is likely to meet the standard exogeneity and relevance criteria (Angrist et al., 1996), as treatment status was (quasi-)random.

2.5. Power

Our main sample consists of 802 households (418 in the treatment group and 384 in the control group) for a period of, on average, 14 months. We used the `pc analytic` command of STATA's `pcpanel` package (Burlig et al., 2017, 2020) to analytically derive the Minimum Detectable Effect Sizes (MDEs). We proceeded as follows. For the ATE we started by estimating Eq. (1) for the sample of control group households. Using the values of the error terms (ϵ_{it}) thus derived, we estimated the error component's autoregressive coefficient assuming an AR(1) process. Assuming that all 802 households' pre- and post-interventions periods were equal to their median lengths (of 94 and 264 days, respectively), we calculated the MDEs for gas and electricity assuming conventional levels for power and statistical significance (0.80 and 0.05, respectively). For the ATE estimates we find an MDE of 0.073 m^3 for gas and of 0.143 kWh for electricity, and a similar analysis for the ATT estimates yields MDEs of 0.106 m^3 and 0.208 kWh for gas and electricity, respectively. These numbers translate into minimum detectable reductions of about 1.8% for electricity and 2.7% for gas. These MDEs are appropriate because with these savings the IHD's payback period would be about 3.5 years while households in the Netherlands are willing to invest in energy-saving options with payback periods of up to 5 years.[8] With more units we would be powered to detect smaller effect sizes, but this would not yield any additional actionable policy advice because with smaller savings the IHD would fail to be cost-effective; see also McCloskey and Ziliak (1996) and McKerracher and Torriti (2013) for a similar reasoning.

[8] The acceptable payback period of five years has been documented by an environmental NGO in the Netherlands; see https://www.milieucentraal.nl/energie-besparen/energiesubsidies-en-leningen/rendement-energiebesparing/.

3. The IHD's impact on energy consumption

3.1. Main results

The overall impact estimates of the IHD are presented in Tables 3 and 4. Table 3 presents the ATE estimates, while the ATT estimates are presented in Table 4. Columns (1) and (2) in each of these two tables present the relevant impact estimates on the daily consumption of, respectively, electricity and gas, while the impacts on the household's total amount of energy consumed per day are presented in column (3) in each of these two tables. We find that our results crucially depend on (how to deal with) the role of space. More specifically, both the size of the impact estimates as well as their standard errors are affected by two modeling decisions: (i) whether the time fixed effects are sample-wide or whether they are allowed to differ between locations (i.e, whether we use calendar day or region-calendar day fixed effects), and (ii) whether standard errors are clustered at either the household level in all locations, or whether they are clustered at the sublocation level (i.e., the two-digit zip code level) in those locations with ISKRA-metered households whose meter could not be updated to become IHD-compatible. As stated in Section 2.4, ex ante the latter would take into account the quasi-randomness of the allocation to treatment and control in those locations where households with unupdatable ISKRA meters were present. Each of these two tables thus consists of four different panels which show the consequences for (the significance of) the treatment estimates of the above two modeling decisions.

Comparing the outcomes of Panels A to those in Panels B in each of these two tables, we find that when imposing the time fixed effects to be identical for all locations, the level of clustering has a differential impact on the size of the standard errors for the impact estimates for gas, electricity and total energy consumption. Taking into account that we were not able to fully randomize all households in each location results in standard errors being smaller for electricity and larger for both gas and the total amount of energy consumed. And comparing the outcomes in Panels A and B to those in Panels C and D in each of the two tables, allowing the time fixed effects to differ between locations has a non-negligible impact on the size of the treatment estimates, even though we already controlled for differential weather patterns in all regressions (by including location-specific and time-varying weather controls). Finally, whereas taking into consideration the quasi-randomness of the smart-meter roll-out has an ambiguous impact on the size of the standard errors when imposing uniform time effects between locations (compare Panels A and B), it results in standard errors that are unambiguously smaller when allowing the time effects to be location-specific (albeit more so for electricity than for gas and total energy); compare Panels C and D in each of the two tables. The spatial component is thus found to crucially affect both the size and significance of the treatment estimates. Imposing the time fixed effects to be identical across locations seems overly restrictive, and hence we regard the outcomes presented in Panels C and D of each of the two tables as the – if anything, conservative – estimates of the IHD's treatment impact.

As shown in column (1) of Panels C and D of Table 3, the Average Treatment Effect of offering the IHD decreases the amount of electricity consumed by 135 Wh per day. Compared with the average amount of electricity consumed by the control group during the intervention period (of 7,661 Wh per day), this translates into a relatively small reduction in electricity consumption of 1.76% which is significant only when the quasi-randomness of smart-meter roll-out is incorporated in the clustering decision (see Panel D). Next, for natural gas, the reduction in the daily amount consumed equals 0.150 m^3 (see column (2) in Panels C and D of Table 3), which translates into a 5.63% reduction in natural gas consumption compared to the control group's mean daily consumption of 2.665 m^3 during the intervention period. This reduction is significant independent of whether standard errors are clustered to reflect the randomization process, or whether they are simply clustered at the household level. The reduction of the total amount of energy consumed by treatment households compared to those in control is essentially a weighted average of the savings for electricity and gas. Upon converting the energy content (measured in Wh) of the type of natural gas supplied in the Netherlands (where 1 m^3 of natural gas equals 9,769 Wh), we find that the amount of energy consumed per day decreases by about 1,580 Wh per day, and this reduction is significant independent of how the standard errors are clustered; see column (3) of Panels C and D in Table 3. Comparing these savings to the average amount of (the specific type of) energy consumed by the control group during the intervention period, we find that the IHD resulted in a 4.70% reduction in the total amount of energy consumed.

Not surprisingly, the ATT estimates presented in Table 4 closely resemble those of the ATE estimates as presented in Table 3 in terms of the consequences of the two modeling decisions discussed above. As explained in Section 2.4, the ATT estimates are obtained using a 2SLS model to correct for the share of treatment households that ended up not receiving an IHD — most often because of technical difficulties or because they lost interest; see Sections 2.2 and 2.4. Again, the more conservative estimates are those presented in Panels C and D of Table 4. We find that the treatment-on-the-treated estimates are 2.17% savings for electricity (which fails to be significant unless the clustering takes into account the imperfect randomization process), 6.94% for gas, and hence 5.81% overall. With oversubscription designs the ATT is typically deemed to be more relevant than the ATE, both academically and from a policy perspective. After all, the cost of the IHD is typically much higher than the marginal costs of recruitment and IHDs only need to be distributed to those who actually want to receive one. Based on the outcomes of Panels C and D of Table 4 we thus conclude that the IHD results in a relatively small (and possibly insignificant) reduction in electricity consumption of 2.2% and in a significant reduction in the amount of gas consumed by 6.9%, resulting in a significant reduction of the total amount of energy consumed of 5.8%.

The 2.2%, savings on electricity that we document are significant, but relatively small — they are even at the lower end of the range of savings of between 2% and 5% as typically achieved by feedback interventions; see Degen et al. (2013), McKerracher and Torriti (2013), Khanna et al. (2021). This may be due to the fact that our study is more high-powered than is typically the case in this literature; as such it is in line with the finding by Karlin et al. (2015) that the documented effect sizes are typically smaller the larger the sample size. With 6.9% the savings on gas are, however, substantially larger — so large that the overall

Table 3
Average Treatment Effects (ATE) of the impact of the IHD on the daily amount of gas and electricity consumed (in m^3 and in Wh, respectively) as well as on the total amount of energy consumed per day (in Wh).

	(1)	(2)	(3)
	Electricity (Wh)	Gas (m^3)	Total energy (Wh)
Panel A			
Treatment	−246.203**	−0.240***	−2572.521***
	(97.650)	(0.069)	(684.859)
Time fixed effects	Day	Day	Day
Clustering	hh	hh	hh
F-value	29.702	33.328	33.333
Panel B			
Treatment	−246.203***	−0.240***	−2572.521***
	(82.102)	(0.072)	(721.735)
Time fixed effects	Day	Day	Day
Clustering	2-digit zip code/ hh	2-digit zip code/ hh	2-digit zip code/ hh
F-value	14.461	31.827	16.273
Panel C			
Treatment	−134.703	−0.150**	−1579.503**
	(95.844)	(0.068)	(684.755)
Time fixed effects	Day × region	Day × region	Day × region
Clustering	hh	hh	hh
F-value	21.146	4.436	7.629
Panel D			
Treatment	−134.703**	−0.150**	−1579.503**
	(61.488)	(0.061)	(615.124)
Time fixed effects	Day × region	Day × region	Day × region
Clustering	2-digit zip code/ hh	2-digit zip code/ hh	2-digit zip code/ hh
F-value	12.739	11.154	8.824
Control group mean	7660.794	2.665	33,597.507
Number of hhs (N)	802	802	802
Observations (N × T)	346,321	343,375	341,771
Household fixed effects	Y	Y	Y
Weather controls	Y	Y	Y

Notes: ATE estimates of the IHD's impact on electricity consumption (in Wh, in column (1)), gas consumption (in m^3, in column (2)) and on the total amount of energy consumed by the household (in Wh, in column (3)). The total amount of energy consumed is calculated using the Wh content of natural gas in the Netherlands, where the energy content of 1 m^3 of natural gas equals 9,769 Wh. Panel A presents the results using sample-wide calendar day fixed effects and with standard errors clustered at the household level, Panel B presents those with sample-wide calendar day fixed effects and with standard errors clustered either at the 2-digit zipcode level (in those regions with non-updatable ISKRA meters) or at the household level, Panel C shows the results using region-calendar day fixed effects and with standard errors clustered at the household level and Panel D shows the results using region-calendar day fixed effects and with standard errors clustered at either the 2-digit zipcode level (in those regions with non-updatable ISKRA meters) or the household level. *, **, and *** indicate significance at, respectively, the 10, 5, and 1 percent critical level. The control group mean is the average amount of energy consumed by the control group households since the start of the intervention, as defined by the median installation date of the IHD in a batch.

savings in household energy consumption are at the top end of the range of savings. As documented by Gerster et al. (2021), providing appliance-specific feedback increases overall savings on electricity consumption by an additional 5 percentage points above and beyond what is achieved by providing information about household-level electricity consumption. We find a similar impact when providing aggregate feedback separately for gas and electricity consumption, possibly because the combination of (i) feedback being provided real-time and (ii) gas being used for just three purposes (cooking, and space and water heating) makes it relatively straightforward for households to assess the energy intensity (and hence also the savings potential) for three of the most energy-intense activities in the household. As such, our study complements the one by Gerster et al. (2021).

3.2. Robustness of the results

As explained in more detail in Section 2.2, a number of implementation issues made our impact evaluation non-standard. We thus probe the reliability of the results presented in Tables 3 and 4 using a series of robustness tests, which are presented in Appendix A.2. First, while the impact of the IHD on gas and total energy is large and significant across all four specifications presented in Tables 3 and 4, the treatment estimate of the impact on electricity is significant only in three of the four. Note that the impact estimate for electricity teetering on the brink of significance is consistent with our analysis having been powered for a minimum detectable effect (MDE) of 0.143 kWh; see Section 2.5. The difference between the actual estimate and the MDE is just 9 Wh, but it does explain why the level of clustering may affect the significance of the estimate. Rather than adjusting clustering we can use randomization inference, a placebo-type approach where units are randomly reassigned treatment status (Fisher, 1925; Barrios et al.,

Table 4
Estimates of the Average Treatment on the Treated (ATT) impact of the IHD on the daily amount of gas and electricity consumed (in m^3 and in Wh, respectively) as well as on the total amount of energy consumed per day (in Wh).

	(1)	(2)	(3)
	Electricity (Wh)	Gas (m^3)	Total energy (Wh)
Panel A			
Instrumented Installation	−300.182**	−0.293***	−3138.556***
	(118.854)	(0.083)	(830.796)
Time fixed effects	Day	Day	Day
Clustering	hh	hh	hh
F-value	29.764	33.316	33.370
Kleibergen–Paap rk LM statistic	389.431	388.532	388.851
Panel B			
Instrumented Installation	−300.182***	−0.293***	−3138.556***
	(100.844)	(0.084)	(846.296)
Time fixed effects	Day	Day	Day
Clustering	2-digit zip code/ hh	2-digit zip code/ hh	2-digit zip code/ hh
F-value	14.982	31.287	16.612
Kleibergen–Paap rk LM statistic	9.847	9.878	9.873
Panel C			
Instrumented Installation	−166.486	−0.185**	−1953.390**
	(118.292)	(0.084)	(844.342)
Time fixed effects	Day × region	Day × region	Day × region
Clustering	hh	hh	hh
F-value	21.150	4.459	7.649
Kleibergen–Paap rk LM statistic	407.406	406.651	407.086
Panel D			
Instrumented Installation	−166.486**	−0.185**	−1953.390***
	(78.230)	(0.074)	(747.816)
Time fixed effects	Day × region	Day × region	Day × region
Clustering	2-digit zip code/ hh	2-digit zip code/ hh	2-digit zip code/ hh
F-value	12.302	11.711	8.998
Kleibergen–Paap rk LM statistic	8.793	8.821	8.816
Control group mean	7660.794	2.665	33,597.507
Number of hhs (N)	802	802	802
Observations (N × T)	346,321	343,375	341,771
Household fixed effects	Y	Y	Y
Weather controls	Y	Y	Y

Notes: ATT estimates of the IHD's impact on electricity consumption (in Wh, in column (1)), gas consumption (in m^3, in column (2)) and on the total amount of energy consumed by the household (in Wh, in column (3)). The total amount of energy consumed is calculated using the Wh content of natural gas in the Netherlands, where the energy content of 1 m^3 of natural gas equals 9,769 Wh. Panel A presents the results using sample-wide calendar day fixed effects and with standard errors clustered at the household level, Panel B presents those with sample-wide calendar day fixed effects and with standard errors clustered either at the 2-digit zipcode level (in those regions with non-updatable ISKRA meters) or at the household level, Panel C shows the results using region-calendar day fixed effects and with standard errors clustered at the household level and Panel D shows the results using region-calendar day fixed effects and with standard errors clustered at either the 2-digit zipcode level (in those regions with non-updatable ISKRA meters) or the household level. *, **, and *** indicate significance at, respectively, the 10, 5, and 1 percent critical level. The control group mean is the average amount of energy consumed by the control group households since the start of the intervention, as defined by the median installation date of the IHD in a batch. The Kleibergen–Paap statistics in columns (1)–(3) are significantly different from zero ($p < 0.01$), indicating that our estimates do not suffer from weak identification bias.

2012); the relevant test statistic is then the share of placebo test runs yielding a treatment estimate that is larger than the estimates we presented in Panels C and D of Table 3. As shown in Table A.2 in Appendix A.2.1, the 95% confidence interval of the share of placebo test runs yielding a treatment estimate larger than the actual ones lies below 0.01 for the consumption of not just natural gas and total energy, but for electricity as well. We thus conclude that our treatment estimates are robust to applying randomization inference, and that the savings on electricity are relatively small yet statistically significant at the 5% level (in line with the results presented in Panel D of Table 3).

Second, we tested whether the results in Table 3 remain unchanged if we restrict our analysis to the sample of households that could be randomized by the researchers – i.e., just those with IHD-compatible smart meters. Doing so reduces the number of households included in the analysis from 802 to 645. Balance between the treatment and control group should be decent – after all, all households with IHD-compatible smart meters were randomized in treatment and control, albeit not with a 50 chance of ending up in either group. But because the probability with which households were assigned to treatment and control within a batch is known, we weighted all observations with the inverse of the probability that a household would be assigned to treatment, or control. Note that this probability is batch-specific, as our randomization rule was such that the probability of an IHD-compatible household ending up in treatment was higher the larger the share of ISKRA-metered households in the batch; see Section 2.3 and Table 1. As

shown in Table A.3 in Appendix A.2.2, the balance of the subsample of IHD-compatible households remains decent when applying inverse probability weighting, and also the regression results are by and large unaffected when restricting the sample to just those households with IHD-compatible smart meters; see Table A.4 in Appendix A.2.2.

Third, even though the sample of households we used in the main analysis is well-balanced (see Table 2), we tested whether our results are robust to explicitly conditioning our impact estimates on the household characteristics we have data on — the one variable that was found to be unbalanced in Table 2, but also when conditioning on all other characteristics as well. As shown in Table A.5 in Appendix A.2.3, the ATE impact estimates using a Random Effects model are very similar to those presented in Panel D of Table 3.

Fourth, we tested whether our results are robust to alternative ways of dealing with outliers. As stated in Section 2.3, eight households had annual electricity consumption levels of 9000 kWh or more, which is well above the 95th percentile of the group of households with the highest energy consumption in the Netherlands (those consisting of five or more household members living in large detached houses). Excluding these eight households from the analysis is warranted if being a very large energy user is better viewed as a household characteristic (with unobserved underlying causes) than as a status that can differ from day to day depending on unobserved idiosyncratic shocks. In any case, our results are robust to winsorizing the amounts of energy consumed at the 0.5% and 1% level (when winsorizing at the top, but also with symmetric winsorization); see Table A.6 in Appendix A.2.4 and also Table B5 in Appendix B (Supplementary Materials).

Fifth, the definition of the start of the intervention period is quite natural for the households in the treatment group that received an IHD — the day on which the IHD was installed. The definition is less clear-cut for those households in the treatment group in whose homes the IHD was not installed — because of technical difficulties, or because of a loss of interest in receiving one; see Section 2.2. In our main analyses (as presented in Tables 3 and 4) we used the median date of IHD installation as the definition of the start of the intervention period for these "non-compliers"; in Appendix A.2.5 we tested whether our results are robust to redefining the start of the intervention period to the day at which the *first* household in a batch received their IHD. The results are presented in Table A.7 in Appendix A.2.5; again, the outcomes are very similar to those presented in Panel D of Tables 3 and 4.

Sixth, we explored the relevance of estimating the amount of electricity produced by those households with solar panels. The share of households owning solar panels is similar between treatment and control (see Table 2), and hence the solar panels' electricity-production capacity is very likely to be similar too. If so, correcting for self-regeneration is then expected to reduce measurement error, but the average treatment effect should remain unchanged. Table A.8 in Appendix A.2.6 presents the results of re-running the regression presented in Panel D of Tables 3 and 4 using the net amount of electricity purchased — that is, without adding the correction for self-consumption of electricity by households that own photovoltaic cells. We find that in percentage terms the impact estimates remain by and large unchanged as they are 1.63% for the ATE and 2.02% for the ATT (compare with 1.76% and 2.17% in Table 3). But we also find that the treatment estimates just fail to be significant, as the respective p-values are 0.120 and 0.117. Correcting for own electricity generation thus increased the precision of the estimates — not the actual size of the treatment effects.

Seventh, we explored whether there are systematic differences in the IHD's impact among three different dimensions: the amount of energy used at baseline, home ownership, and also outdoor temperatures. The higher the baseline energy consumption, the easier it may be to save, and hence the more effective the feedback is expected to be (see for example Allcott, 2011b; Andor et al., 2020; Asensio and Delmas, 2015; Tiefenbeck et al., 2018; Gerster et al., 2021). Similar considerations are likely to apply to the use of natural gas for heating in Winter and the use of electricity for cooling in Summer (Allcott and Rogers, 2014; Frederiks et al., 2015) – although in the Netherlands the Summer effect for electricity is likely to be less pronounced because air conditioning was, at the time of our study, not very widespread yet.[9] Also, characteristics like home ownership are likely to matter because of the split incentives issue among landlords and renters (as opposed to owner-occupied homes see; Frederiks et al., 2015; Aydin et al., 2017) and also because of the indirect determinants of baseline consumption levels like income (Abrahamse et al., 2007; Costa and Kahn, 2013; Podgornik et al., 2016). In contrast to the above findings we find that, when allowing for type-specific time trends, We cannot reject the null of equal effectiveness in reducing energy consumption for each of these three subgroups; see Table A.9 in Appendix A.2.7. Having said that, the results provide weak evidence that the treatment effect on electricity consumption is larger for households with above-median electricity consumption at baseline; see the results in column (1) of Panel A of Table A.9. Indeed, the difference in the impact estimates between the above- and below-median electricity consumers is large, but the coefficients are not estimated with sufficient precision for this difference to be significant. Still, the suggestive evidence that the IHD is more effective in inducing electricity savings among the above-median intensive energy uses echoes the results found by, among others, Allcott (2011b), Byrne et al. (2018), Tiefenbeck et al. (2018) and Andor et al. (2020).

The results we presented in Tables 3 and 4 are thus robust, qualitatively but also quantitatively, against all seven robustness tests. We thus conclude that the successful installation of the IHD among those households interested in receiving the IHD (i.e., the ATT) induces savings of 6.9%, 2.2% and 5.8% on the amount of, respectively, gas, electricity and total energy consumed.

[9] According to TNO (2021), an independent not-for-profit research organization specializing in innovation and technological change, about 6% of the households in the Netherlands owned an air conditioner in 2018 (either mobile, or fixed); see p. 8 of https://publications.tno.nl/publication/34639092/k8jFX5/TNO-2021-R12657.pdf.

4. Mechanisms: knowledge, behavior and attitudes

We thus find that the IHD induces reductions in the amounts of electricity and gas consumed (albeit more prominently so for the latter), and also in the total amount of energy consumed; see Panel D in Tables 3 and 4. As shown in Table A.9 in Appendix A.2.7, we found little evidence for the effectiveness of the IHD varying with household characteristics such as baseline energy consumption, home ownership status or outdoor temperatures — except, possibly, for electricity, where the savings seem to be driven by substantial reductions in consumption by households that were consuming above-median amounts of electricity at baseline. To gain insight into the possible underlying mechanisms, all households were invited, via email, to fill out an internet survey. Each household was invited to participate just once, but because of the ongoing treatment roll-out we sent out the survey at four different moments to ensure an approximately equal time lapse between the start of the intervention and the survey for each household. Survey participation was incentivized by means of a lottery in which those who filled out the survey had the chance to win one of twenty gift cards, each valued at 50 euros. We ended up with an overall response rate of 55% (442 of the 802 households in our main sample filled out the survey), but the share of respondents among the treatment households was higher than that among the control group households: 62 percent of the treatment group filled out the survey, versus 48 percent of the control group ($p < 0.01$).

We use the survey responses to analyze (i) how the energy savings were achieved (in particular, by changing behavior, or rather via the purchase of more energy-efficient appliances and technologies), (ii) whether the IHD improved households' knowledge and understanding of their energy usage (either in terms of what behaviors and uses are most energy-intensive, in their knowledge of the overall costliness of using energy, or their belief that they can effectively influence their energy consumption), and (iii) whether it affected the households' attitudes towards the importance of saving energy and, more general, towards taking climate action.

The reliability of the assessment depends on whether there was non-random selection into survey participation, and a more detailed analysis of how households used the IHD is helpful in interpreting the underlying mechanisms of the observed reductions in energy consumption. We thus first test, in Section 4.1, whether the sample of respondents is similar to the sample of RCT participants and whether respondent characteristics are balanced between treatment and control, and in Section 4.2 we analyze how treatment households made use of the IHD. The results of the analysis of the potential mechanisms giving rise to the observed savings in energy are presented in Section 4.3.

4.1. Representativeness of the survey respondents, and the issue of non-response bias

Overall, the response rate to the survey was 55%. Despite this relatively high response rate (Holtom et al., 2022), non-respondent bias may be a concern (Rindfuss et al., 2015), especially because of the 14 percentage point difference in the response rates among the treatment and control group households. A comparison of the mean characteristics of households that had filled out the survey and those that had not provides insight into the representativeness of the survey respondents vis-à-vis the full sample, and a comparison of the mean characteristics of the survey respondents of the treatment group and of the control group provides insight into the extent to which our treatment effect estimates on the survey responses may be biased by treatment-induced self-selection into the survey. The former bicomparison is important for external validity, the latter is important because it provides insight into the internal validity of our survey outcomes. A difference of 14 percentage points in response rates is sizable and it may have been the result of, for example, positive reciprocity by the treatment group ("you gave us an IHD for research purposes, so it is only polite for us to tell you how we evaluate the experience") or negative reciprocity among the control group households (think of disappointment or spite, for not having received an IHD). Of course, we cannot observe these sentiments or considerations among the non-respondents, but the key question is whether or not these sentiments (or considerations) are orthogonal to all the key factors affecting, say, the scope for saving energy. In other words, our analysis of the survey data is biased if, for example, those control group households with higher baseline energy consumption are more likely to be disappointed by not having won an IHD (and hence refuse to fill out the survey) than those with lower baseline energy consumption. Similarly, our analysis is also biased if the propensity to reciprocate among the treated households with higher baseline energy consumption is different than among those with lower baseline energy consumption.

In Appendix A.3 we present the results the two types of bicomparisons discussed above. As shown in Table A.10, standard bicomparison tests suggest that the respondent households differ from the non-respondent households in, in total, seven characteristics, in four different groups: respondent households tend to be older, they tend to be higher educated (as the highest level of education attained in their household is more likely to be tertiary than in non-respondent households), they are more likely to live in larger homes (i.e., more likely to live in a detached house, and less likely to live in "other" housing), and they are more likely to have solar panels at baseline. These differences are by and large robust to Multiple Hypotheses Testing (see Table B6 in Appendix B, Supplementary Materials), and also the normalized differences presented in Table A.10 suggest that households that responded to the survey are not fully representative of the sample of households that participated in the RCT as quite a few are very close to the threshold of 0.25 standard deviations typically used in this literature, but quite a few of them are close. Fortunately, from an internal validity perspective, the differences between respondent households of the treatment group and those of the control group tend to be (much) smaller; see Table A.11. Among the survey respondents the elderly households are overrepresented in the treatment group, and the same holds for households living in apartments. These differences are, however, not robust to Multiple Hypotheses Testing; see Table B7 in Appendix B (Supplementary Materials).

Self-selection thus seems to be less of a concern for the internal validity of the analysis of the survey than for the overall representativeness of the survey respondents vis-à-vis the overall sample. We test this more formally by using regression analysis to analyze households' propensity to fill out the survey. The regression model is as follows:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{T}\boldsymbol{X}\boldsymbol{\gamma} + \boldsymbol{\epsilon}. \tag{4}$$

Table 5
Analysis of the propensity of filling out the survey, and the role of treatment status therein.

	(1) Characteristic (β)	(2) Char. × treatment (γ)
Treatment		0.095
		(0.114)
1 person hh	0.078	−0.059
	(0.062)	(0.090)
> 2 hh members	0.078	−0.136
	(0.077)	(0.108)
Elderly hh	0.024	0.165*
	(0.064)	(0.088)
Household with children below 18	−0.037	0.075
	(0.063)	(0.086)
Just primary education	−0.184	0.021
	(0.163)	(0.219)
Higher tertiary education	0.141***	0.012
	(0.053)	(0.076)
Apartment	−0.028	0.068
	(0.078)	(0.108)
Corner house	0.096	−0.092
	(0.071)	(0.102)
Semi-detached house	0.029	−0.054
	(0.079)	(0.115)
Detached house	0.150	−0.035
	(0.096)	(0.135)
Dwelling owned by the household	−0.031	0.013
	(0.066)	(0.095)
House built after 1974	−0.032	0.083
	(0.052)	(0.074)
Owned photovoltaic cells at time of randomization	0.124**	−0.045
	(0.052)	(0.075)
Above-median baseline gas consumption	0.007	0.002
	(0.053)	(0.076)
Above-median baseline electricity consumption	−0.022	−0.016
	(0.056)	(0.079)

Note: Household characteristics correlated with the decision to fill out the survey. The first column presents the average marginal effects of selecting into the survey (i.e., the elements of coefficient vector β in Eq. (4)) and the second column presents those of the interaction effect of the characteristic and treatment status (i.e., whether the effects are significantly different for the treatment group households, as captured by vector γ in Eq. (4)). Robust standard errors are presented in parentheses; *, **, and *** indicate significance at, respectively, the 10, 5, and 1 percent critical level. As shown in Tables O8 and O9 in the Online Appendix, none of the differences are robust to Multiple Hypotheses Testing.

In this model, $\boldsymbol{y}$ is a vector of binary values indicating whether a household filled in the survey (with a value of 1 if the household did so, and zero otherwise), $\boldsymbol{X}$ denotes the matrix of household characteristics as observed at baseline and $\boldsymbol{T}$ denotes the vector of treatment statuses of each of the households. We thus fully interacted all covariates (the intercept, as well as all baseline variables) with the treatment dummy. Coefficients in vector $\boldsymbol{\beta}$ reflect how survey respondents differ from survey non-respondents among our sample of participants in the study, and any significant coefficients in vector $\boldsymbol{\gamma}$ then captures treatment-related differences in the propensity to participate in the survey. The results of this probit analysis are presented in Table 5, where the coefficients reflect the average marginal effects.

Regarding Table 5, note that the coefficients were obtained by running the above model (see Eq. (4)), but for expositional succinctness we parsed them into two different columns, with columns (1) and (2) presenting the $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ coefficients, respectively. As can be inferred from column (1) of Table 5, we find that higher educated households were more likely to fill out the survey (i.e., those with higher tertiary education; $p = 0.008$), and the same holds for households having solar panels at baseline ($p = 0.018$). And column (2) shows that elderly households in the treatment group were significantly more likely to fill out the survey than elderly households in the control group ($p = 0.060$). None of the other interaction effects are significantly different from zero at conventional levels (with all $p \geq 0.209$). Also, none of the significant differences survive Multiple Hypothesis Testing; see Tables O8 and O9 for the full set of MHT-adjusted p-values applied to coefficient vectors $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$, respectively.

Even though the differences in survey respondents' characteristics between treatment and control households are small and not robust to MHT, we mitigate the impact of any possible remaining selection bias by using the full set of household and home characteristics (as shown in Table A.11) as controls in all regressions presented in Section 4.3.

4.2. Treatment households' overall appreciation and usage of the IHD

Before turning to the analysis of the mechanism giving rise to the energy savings documented above, we assess the treatment households' appreciation of the IHD. The survey answers suggest that the display was successful in increasing the salience of energy

consumption. Most of the respondents reported that they kept the display in a central place of their home such as the living room (66%) or the kitchen (18%), and 92% of the respondents indicate that the IHD was still in use at the time of the survey. In the first weeks after installation more than 80% of households looked at the IHD's information more than five times a week. While in the week preceding the survey this share had gone down to 52%, overall usage was still relatively high; more than six months after the installation of the IHD, only 14% of the recipient households reported to have ceased looking at the display altogether (cf. Hargreaves et al., 2013). Consistent with this, at the time of the survey about 70% of the households stated that they still find the IHD useful. Finally, regarding the IHD's functionality, 85% of the households stated that they predominantly used the direct feedback (either in physical units, or in costs); less than a quarter used the daily budget function screen on a regular basis.

Of course, these numbers are the average responses of those treatment households that filled out the survey. This may be an overestimate of the actual appreciation among all households in the treatment group if those who were less satisfied were less likely to fill out the survey. The number of treatment households having ceased to look at the IHD six months after installation is between 14% (if the appreciation of the IHD by the respondent households is the same as that of the non-respondents) and 53% (if all 159 non-respondents in the treatment group had ceased looking too).[10] Similarly, the share of households still finding the IHD useful after six months ranges between 43% and 70%.[11]

4.3. Assessing the IHD's impact on behavior, knowledge, and attitudes

With the above caveats about the potential presence of non-response bias as presented in Section 4.1 in mind, Table 6 presents the Average Treatment Effects[12] of the IHD on self-reported energy saving actions (change in behavior, and adoption of energy-saving appliances and technologies; see columns (1) and (2)), on the observed adoption of solar panels (in column (3)), on households' knowledge about the energy intensity of the various types of domestic energy uses, the overall costs of energy consumption, and their self-efficacy with respect to saving energy (in columns (4)-(6)), and also on their attitudes towards energy saving, and more generally, the importance of combating climate change (in columns (7) and (8)). As the survey asked households about their adoption of a fairly large number of energy use behaviors and energy-saving technologies (9 and 13, respectively), we mitigate the issue of Multiple Hypotheses Testing by presenting the results in terms of indices equal to sum of behaviors and technologies adopted (Kling et al., 2007; Anderson, 2008); see Section O.2 in the Online Appendix for details. For the analysis in column (1) we count the number of behaviors households reported to consistently implement, and for column (2) we count the number of energy-efficiency investments households report to have undertaken over the past six months. Households' adoption of solar panels in the intervention period (as analyzed in column (3)) can be monitored by checking whether or not the household starting to sell back electricity to the grid sometime after the roll-out of the IHDs in their batch. To test each household's knowledge of the relative energy intensity of the various energy uses (see column (4)) we asked respondents to rank a number of fairly broad energy consumption categories (heating, lightening, cooking etc.) by their energy intensity; cf. Attari et al. (2010). Every category placed in the correct position yields 5 points; one point is subtracted for every position the household's answer deviated from the correct one. We asked households to rank six categories, and hence the overall score ranges from 0 to 30. For column (5) we asked households to estimate their monthly gas and electricity expenditures, and compared that to their actual energy bill. Self-efficacy and the two types of attitudes (in columns (6)-(8)) were obtained using scales consisting of, respectively, four, six and six items; we again constructed indices thereof to reduce the issue of Multiple Hypotheses Bias. More specifically, for all items within these scales we assigned a value of 1 if the household agreed with the statement, and 0 otherwise; the index for each of the three scales is then equal to the sum of the scores for each item therein.[13] Panel A of Table 6 presents the overall Average Treatment Effects; Panel B tests whether the impact differs between households that were above- or below-median users of energy at baseline.

Regarding the overall impacts (as shown in Panel A of Table 6), we find that the IHD affected households' propensity to adopt energy-saving technologies and solar panels (columns (2) and (3)) as well as their accuracy in assessing the amount they pay per month (column (5)), with conflicting results. Households in the treatment group had adopted slightly fewer new energy-saving technologies (0.339, $p = 0.017$), and they were sightly less likely to adopt solar panels (3.5 percentage points less likely; $p = 0.095$). But we also find that households in the treatment group were 10.6 percentage points less likely to underestimate the amount of money they need to pay for energy per month ($p = 0.013$), suggesting that they are better aware of the costs of energy usage. Given that we find that the IHD resulted in lower consumption of both gas and electricity, the mechanism seems to have been via making energy costs more salient — as evidenced by the treatment households' lower probability of underestimating their energy bill.

[10] Of the 259 treatment households that filled out the survey, 36 (or 14%) stated that they ceased looking at the IHD six months after installation. Assuming that all of the (418 - 259 =) 159 non-respondent treatment households also had ceased using the IHD, the maximum share of non-users is equal to ((418 - 36 - 159)/418 =) 53%.

[11] Of the 259 treatment households that filled out the survey, 181 (or 70%) stated that they still found the IHD useful; this thus provides the upper bound of the range. Assuming that none of the non-respondent treatment households would find the IHD useful, the minimum share of treatment households rating the IHD as useful is equal to (181/418 =) 43%.

[12] For simplicity we present just focus on presenting the ATE results. The ATT results are very similar; results available upon request.

[13] We used 7-point Likert scales (ranging from "completely disagree" via "neutral" to "fully agree") for each of the six items for each of the two attitude categories. Creating an index whether a household agrees (versus being neutral or disagree) rather than the Likert score itself minimizes the scope for interpersonal subjectivity bias.

Table 6
Treatment impacts (main effects, and heterogeneous effects) on energy-saving behavior, knowledge and attitudes.

	Energy saving activities			Knowledge and Perceptions			Attitudes	
	(1) Behavioral change	(2) Technology adoption	(3) Owns solar panels	(4) Knowledge relative energy intensities	(5) Underestimate energy bill	(6) Perceived self-efficacy	(7) Attitude towards saving energy	(8) Attitude towards energy transition
Treatment	0.095	−0.339**	−0.035*	−0.129	−0.106**	−0.087	0.034	0.108
	(0.212)	(0.140)	(0.021)	(0.099)	(0.042)	(0.083)	(0.135)	(0.179)
R^2	0.143	0.212	0.193	0.156	0.088	0.134	0.107	0.180
Treatment ×below-median energy user	0.531***	−0.468**	−0.045*	−0.122	−0.042	−0.049	0.075	0.128
	(0.166)	(0.188)	(0.024)	(0.125)	(0.052)	(0.145)	(0.192)	(0.206)
Treatment ×above-median energy user	−0.352	−0.206	−0.025	−0.136	−0.172***	−0.126	−0.008	0.088
	(0.230)	(0.171)	(0.030)	(0.105)	(0.054)	(0.089)	(0.125)	(0.247)
R^2	0.163	0.213	0.193	0.156	0.094	0.134	0.107	0.180
Observations	442	442	442	442	437	442	442	442
Control group mean	3.541	6.754	0.104	6.780	0.379	2.508	4.514	3.721
Region fixed effects	Y	Y	Y	Y	Y	Y	Y	Y
Survey time fixed effects	Y	Y	Y	Y	Y	Y	Y	Y
HH and home characteristics	Y	Y	Y	Y	Y	Y	Y	Y

Notes: For the (translated version of the) survey questions, see Appendix O.2. Above- and below-median energy users are coded on the basis of users' electricity and gas consumption, weighted by their average prices. Standard errors, presented in parentheses, are clustered either by 2-digit zip code (in those regions with non-updatable ISKRA meters) or at the household level; *, **, and *** indicate significance at, respectively, the 10, 5, and 1 percent critical level. As shown in Table B10 in Appendix B (Supplementary Materials), only two of the main impacts remain (borderline) significant upon Multiple Hypothesis Testing — those for technology adoption and for underestimating the energy bill (in columns (2) and (5), respectively). The difference in technology adoption remains significant at $p = 0.073$ according to Anderson's sharpened q; the adjusted p-values for the other three tests range between 0.112 and 0.203. The difference in the propensity to underestimate the energy bill remains significant at $p = 0.073$ and $p = 0.098$ according to Anderson's sharpened q and Sidak–Holm, and the adjusted p values for the other two tests are 0.102 and 0.193. For all other differences the MHT-corrected p-values are 0.236 higher; see Table B10 in Appendix B (Supplementary Materials).

The conclusion that salience of the costs of using energy is likely to have been the key driver of the decrease in energy consumed is reinforced by the outcomes of a series of Multiple Hypotheses Tests; see Table B10 in Appendix B (Supplementary Materials). Here, only two effects show signs of significance; technology adoption (column (2) in Table 6, with a p-value of 0.073 for Anderson's sharpened q (Anderson, 2008), and with Sidak–Holm (Abdi, 2007), Bonferroni–Holm (as implemented in Jones et al., 2019) and Westfall–Young (Westfall and Young, 1993) reporting p-values of 0.112 or higher) as well as the impact on the propensity not to underestimate the energy bill (column (5) in Table 6, with p-values of 0.073 and 0.098 Anderson sharpened q and Sidak–Holm, and with p-values of 0.102 and higher for Bonferroni–Holm and Westfall–Young. The results on the reduced likelihood of underestimating one's energy bill are thus slightly more robust to MHT than the results on technology adoption. And given that we document a decrease in energy consumption in Section 3, cost salience seems to be the main mechanism causing consumption to decrease.

These result may, however, hide important differences in the IHD's impact between the bigger and the smaller energy users. While not significant, the results presented in Table A.9 in Appendix A.2.7 suggest that it was especially the above-median consumers of electricity that realized the largest savings. As shown in column (1) of Panel B of Table 6, we find a marked heterogeneous treatment effect on the number of energy-saving behaviors households implemented, with the IHD having resulted in a marked increase in these behaviors among the below-median energy users. For these users the beneficial impact of implementing more energy-saving behaviors is at least partly offset by a slightly lower propensity to adopt new technologies — both stated as well as revealed (as shown in columns (2) and (3) of Panel B, respectively). For the above-median energy users the only variable that shows up significantly is the reduction in the propensity to underestimate one's energy bill. Given that we found suggestive evidence that it was the above-median energy users for whom the IHD generated the largest reductions in electricity consumption (see Table A.9), we conclude that for these households the reduction in energy consumption was driven by the increased awareness of the costs of consuming energy (see column 5)).

None of the other impacts, neither the overall impacts nor the heterogeneous impacts, are significantly different from zero. We find no significant differences in households' knowledge of the energy-intensity of various appliances and activities (see column (4)) or in households' self-efficacy (see column (6)), and the same holds for both the differences in households' energy attitudes towards saving energy (column (7)) and in their overall support for the climate action via the energy transition (in column (8)).

5. Conclusion

We implemented an RCT to estimate the impact of real-time feedback on residential energy consumption. The feedback was provided by means of an energy in-home display for gas and electricity, separately. We recruited households interested in receiving a display that provides insight into metered natural gas and electricity consumption in seven regions in the Netherlands. We randomly offered a display to half of the applicant households; the other half did not receive such offer. We monitored all households' energy consumption for a period of, on average, ten months post-treatment.

We find that the energy display led to a reduction of daily gas and electricity consumption of 6.9 and 2.2 percent, respectively, and hence in an overall reduction in total residential energy consumption of 5.8 percent. Feedback on the aggregate amount of electricity consumed typically results in savings of between 2% and 5% (Degen et al., 2013; McKerracher and Torriti, 2013; Khanna et al., 2021). Our impact on electricity consumption is thus at the lower end of the range, and this is in line with the finding that the impact of feedback is typically lower in more high-powered studies; see Karlin et al. (2015). However, we also find that providing real-time feedback on the aggregate amount of gas consumed results in savings of 6.9% – much higher than the savings on electricity, and also higher than typically reported in the literature (Tiefenbeck et al., 2018). Establishing the link between the use of an appliance or technology and its consequences for energy consumption is easier the more targeted the feedback is. While many different appliances are powered by electricity, the number of activities powered by gas is fairly limited. As such, our findings complement those by Gerster et al. (2021), who find that providing appliance-specific feedback increases induced electricity savings by an additional 5 percentage points. Our estimates are in line with this finding.

Regarding the underlying mechanisms, survey respondents indicate that they frequently consulted the device for information; based on the survey evidence, we estimate that six months after the IHD had been installed more than half of the participants were still actively using it. Savings seem to have been driven especially by increased awareness of the costliness of consuming energy, as treatment households are significantly less likely to underestimate the size of the energy bill than their peers in the control group. We thus conclude that the display seems to have been effective in helping households to better align their actual energy consumption with their privately optimal consumption.

The findings of this study contribute to the recent literature that stresses the importance of concrete and relevant information to optimize the effect of energy consumption feedback. This may involve providing feedback in a real-time manner (Darby, 2006) as well as offering concrete energy saving information advice (see e.g. Abrahamse et al., 2005; Gerster et al., 2021; Schleich et al., 2017). The IHD we tested in this study combined both features, and its effectiveness in reducing energy consumption thus provides support for both these claims.

CRediT authorship contribution statement

Mirthe Boomsma: Writing – original draft, Methodology, Formal analysis, Data curation, Conceptualization. **Kees Vringer:** Writing – review & editing, Writing – original draft, Data curation, Conceptualization. **Daan van Soest:** Writing – review & editing, Writing – original draft, Supervision, Investigation, Formal analysis, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

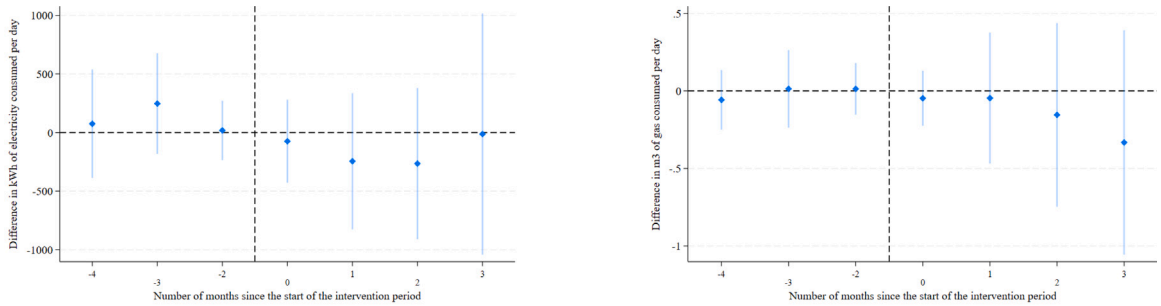
Appendix A

A.1. Testing the quasi-randomness of ISKRA roll-out

We provide two pieces of evidence supporting the claim that the roll-out of the ISKRA meters had been quasi-random, and hence that the amount of energy consumed by ISKRA-metered households (together with the energy consumption by households with IHD-compatible smart meters that were randomized into the control group) can serve as a counterfactual for the amount of energy consumed by the treatment households. First, we checked whether households with ISKRA meters and those with IHD-compatible meters are similar in terms of their characteristics; see Table A.1. Second, we analyze how the amounts of natural gas and electricity consumed by ISKRA households compares to that of the non-treated and the yet-to-be-treated households with an IHD-compatible smart meter (i.e., in the pre-intervention period); see Fig. A.1. We briefly discuss these in turn.

Table A.1 presents the differences between households with IHD-compatible and those with ISKRA (i.e., IHD-incompatible) meters, in those locations in which the ISKRA-update could not be installed. We find that households with unupdatable ISKRA meters were very similar to those with IHD-compatible meters. Only one of the twenty house- and household characteristics shows up significantly (ISKRA meters are more prevalent in terraced houses; $p = 0.073$), and we also find that this difference fails to be robust to multiple hypotheses testing (MHT); see Table B2 in Appendix B (Supplementary Materials). As shown in that table, the MHT-corrected p-values are 0.781 or higher, for all characteristics and across a variety of MHT tests. Also, none of the normalized differences as presented in Table A.1 are larger than 0.25 standard deviations, the critical imbalance threshold typically applied in these types of analyses (Imbens and Wooldridge, 2009; Imbens and Rubin, 2015).

Similar characteristics do not necessarily imply identical (temporal) patterns of energy consumption, and hence Fig. A.1 presents the graphs depicting the differences in the amount of gas and electricity consumed by households with unupdatable (and hence IHD-incompatible) ISKRA meters and by those with IHD-compatible smart meters (those in the control group, and those in the treatment group prior to the start of the intervention). Using month 0 to denote the month in which the median treatment household in a batch received their IHD, we normalized the eventtime graphs at month −1 and focus on the period ranging from four months before to four months after the median installation date for each batch. As is clear from the two panels in Fig. A.1, the amounts of energy consumed by households with unupdatable ISKRA meters were very similar to those with IHD-compatible meters in terms of both their gas and electricity consumption before and after the start of the intervention period. These results reinforce the claim that it was quasi-random which households ended up with which smart meter type: ISKRA and IHD-compatible households are not only very similar in terms of their characteristics (see Table A.1), but also in terms of their gas and electricity consumption (see Fig. A.1).



(a) Difference in daily amount of electricity consumed

(b) Difference in daily amount of gas consumed

Fig. A.1. The differences in the consumption of electricity and natural gas between households with (unupdatable) ISKRA meters and the (non-treated and yet-to-be-treated) households with IHD-compatible smart meters, normalized around the month prior to the one of the median installation date of the IHD in each batch.

Note: Difference in the daily amount of electricity and natural gas consumed, in respectively the left- and right-hand side graph, by ISKRA-metered households (which did not receive an IHD, by definition) and those of households with IHD-compatible smart meters (those in the control group, as well as the pre-intervention consumption data of those in the treatment group), for the period of four months before to four months after the start of the intervention. The start of the intervention, month 0 in these graphs, is defined as the month in which the median treatment household in a batch received their IHD. The difference-in-difference is calibrated by leaving out the dummy for the month prior to the start of the intervention, month -1; see Sun and Abraham (2021). The vertical bars denote the 95% confidence intervals.

Table A.1
Comparison of the characteristics of applicant households with and without IHD-compatible smart meters, in the six regions in which the ISKRA-update could not be installed.

Variable	(1)		(2)		(3)		(4)	(5)
	IHD-compatible hhs		ISKRA-metered hhs		Total		Difference	Normalized diff.
	N	Mean/SE	N	Mean/SE	N	Mean/SE	(1)–(2)	(1)–(2)
1 person hh	464	0.269	151	0.291	615	0.275	−0.022	−0.049
		(0.021)		(0.037)		(0.018)		
2 person hh	464	0.595	151	0.576	615	0.590	0.019	0.038
		(0.023)		(0.040)		(0.020)		
> 2 hh members	464	0.136	151	0.132	615	0.135	0.003	0.010
		(0.016)		(0.028)		(0.014)		
Elderly hh	464	0.248	151	0.219	615	0.241	0.029	0.068
		(0.020)		(0.034)		(0.017)		
Household with children below 18	465	0.366	151	0.371	616	0.367	−0.005	−0.011
		(0.022)		(0.039)		(0.019)		
Just primary education	464	0.017	151	0.026	615	0.020	−0.009	−0.067
		(0.006)		(0.013)		(0.006)		
Secondary or lower tertiary education	464	0.481	151	0.404	615	0.462	0.077	0.154
		(0.023)		(0.040)		(0.020)		
Higher tertiary education	464	0.502	151	0.570	615	0.519	−0.067	−0.135
		(0.023)		(0.040)		(0.020)		
Apartment	464	0.241	151	0.278	615	0.250	−0.037	−0.085
		(0.020)		(0.037)		(0.017)		
Terraced house	464	0.235	151	0.166	615	0.218	0.069*	0.168
		(0.020)		(0.030)		(0.017)		
Corner house	464	0.155	151	0.159	615	0.156	−0.004	−0.010
		(0.017)		(0.030)		(0.015)		
Semi-detached house	464	0.194	151	0.219	615	0.200	−0.025	−0.061
		(0.018)		(0.034)		(0.016)		
Detached house	464	0.151	151	0.166	615	0.154	−0.015	−0.041
		(0.017)		(0.030)		(0.015)		
Other (farmhouse, houseboat)	464	0.024	151	0.013	615	0.021	0.010	0.073
		(0.007)		(0.009)		(0.006)		
Dwelling owned by the household	464	0.573	151	0.576	615	0.574	−0.003	−0.006
		(0.023)		(0.040)		(0.020)		
House built after 1974	464	0.403	151	0.331	615	0.385	0.072	0.148
		(0.023)		(0.038)		(0.020)		
Owned photovoltaic cells at time of randomization	465	0.396	151	0.344	616	0.383	0.051	0.105
		(0.023)		(0.039)		(0.020)		
Baseline average daily gas consumption (m^3)	465	1.925	151	1.990	616	1.941	−0.065	−0.029
		(0.105)		(0.190)		(0.092)		
Baseline average daily electricity consumption (kWh)	465	8.006	151	7.992	616	8.003	0.014	0.003
		(0.210)		(0.397)		(0.186)		
Baseline average daily energy expenditures, in Euros	465	2.852	151	2.892	616	2.862	−0.040	−0.020
		(0.090)		(0.171)		(0.080)		

Notes: Sample statistics (means, and standard errors) of the applicant households with IHD-compatible smart meters (in column (1)) and of those with IHD-incompatible (ISKRA) smart meters (in column (2)), as well as their absolute and normalized differences (in columns (4) and (5), respectively), in the six regions in which the ISKRA-update could not be installed; column (3) presents the means and standard errors of the overall sample. The table was constructed using the iebaltab command of the STATA's ietoolkit package (DIME Analytics, 2019). *, **, and *** indicate that the standard *t*-test of the difference presented in column (4) is significant at, respectively, the 10, 5, and 1 percent critical level. Only the share of households living in terraced houses differs significantly (at $p < 0.10$) between the two subsamples; when correcting for Multiple Hypotheses Testing (MHT) all absolute differences are insignificant (with MHT-corrected *p*-values being equal to 0.781, or higher) according to a series of MHT tests (Anderson sharpened *q* (Anderson, 2008), Bonferroni–Holm (as implemented in Jones et al., 2019), Sidak–Holm (Abdi, 2007), and Westfall–Young (Westfall and Young, 1993); see Table B2 in Appendix B (Supplementary Materials).

A.2. Robustness tests

A.2.1. Clustering and randomization inference

While we were able to update the ISKRA meters in one location (Location 5; see also Table 1), we were not able to do so in the other six locations. Because we could not randomly allocate all households to treatment in control in those locations, we cluster the standard errors at the 2-digit zip code level in those areas where one or more households were endowed with an ISKRA meter. Alternatively, we could have used randomization inference, a placebo-type approach where units are randomly reassigned treatment status; see for example Barrios et al. (2012). We applied randomization inference using Stata's ritest command (Hess, 2017), with 1000 replications for each regression; the "as-if" *p*-value is then the share of the placebo runs that yield a placebo treatment estimate that is larger than the actual ones reported in Panel D of Table 3. The results are presented in Table A.2. As is clear from this table, the results presented in Table 3 are robust to randomization inference; the 95% confidence interval of the share of placebo tests yielding a treatment estimate that is larger than our estimate lies below 0.004 for both electricity and gas, and the same holds also for the total amount of energy consumed.

A.2.2. Robustness to dropping all households with IHD-incompatible smart meters

We test the robustness of our results to using just the sample of households with IHD-compatible smart meters — those ISKRA-metered households for which we were unable to install the update. As explained in Section 2.2 we implemented the project under the assumption that the allocation of IHD-compatible meters had been quasi-random; to maintain an equal allocation to each of the two experimental groups, the larger the share of unupdatable ISKRA-metered households in a batch, the larger the share of IHD-compatible households allocated to treatment; see Section 2.3. Overall, 65% of the households with IHD-compatible smart meters were randomly allocated to treatment, and 35% to control.

Table A.2
Outcomes of the randomization inference tests for the results presented in Panel D of Table 3.

	(1) Mean	(2) quasi p-value	(3) 95% CI
Impact on electricity consumption (Wh)	−134.7	0.000	[0, 0.004]
Impact on gas consumption (m3)	−0.150	0.000	[0, 0.004]
Impact on total energy consumption (Wh)	−1579.5	0.000	[0, 0.004]

Notes: Quasi p-values obtained using randomization inference tests, based on 1000 replications using Stata's ritest command Hess (2017). The mean values in column (1) replicate the treatment impact estimates of the original regressions, column (2) presents the quasi-p-value of the randomization inference test (i.e., the share of placebo runs yielding treatment estimates that are larger than the ones using the actual treatment allocation), and column (3) shows the 95% confidence interval of the quasi-p-value.

Table A.3
Inverse probability weighted balance table for the IHD-compatible households in treatment and control.

	Diff. in means using all hhs			IPW-weighted diff. in means using IHD hhs		
	(1) Difference in means	(2) Std. error	(3) Regular p-value	(4) Difference in means	(5) Std. error	(6) Regular p-value
1 person hh	0.030	0.032	0.357	−0.001	0.040	0.975
2 person hh	−0.028	0.035	0.421	0.001	0.045	0.986
> 2 hh members	−0.002	0.024	0.944	0.000	0.028	0.987
Elderly hh	−0.063	0.031	0.042	−0.061	0.039	0.121
Household with children below 18	0.036	0.034	0.278	0.080	0.047	0.090
Just primary education	−0.000	0.011	0.976	−0.009	0.012	0.455
Secondary or lower tertiary education	0.009	0.035	0.806	0.016	0.048	0.742
Higher tertiary education	−0.008	0.035	0.813	−0.007	0.048	0.891
Apartment	−0.040	0.031	0.194	−0.091	0.032	0.005
Terraced house	0.042	0.031	0.181	0.080	0.042	0.054
Corner house	0.011	0.028	0.701	0.003	0.035	0.939
Semi-detached house	0.006	0.026	0.814	0.011	0.041	0.781
Detached house	−0.013	0.023	0.579	−0.009	0.037	0.807
Other (farmhouse, houseboat)	−0.006	0.009	0.491	0.005	0.013	0.708
Dwelling owned by the household	−0.011	0.035	0.751	0.035	0.048	0.462
House built after 1974	−0.018	0.034	0.597	−0.023	0.044	0.600
Owned photovoltaic cells at time of randomization	−0.030	0.035	0.385	0.003	0.047	0.953
Baseline average daily gas consumption (m3)	−0.092	0.147	0.529	0.059	0.269	0.828
Baseline average daily electricity consumption (kWh)	−0.107	0.313	0.733	0.303	0.421	0.473
Baseline average daily energy expenditures, in Euros	−0.081	0.130	0.533	0.099	0.234	0.674

Notes: columns (1)–(3) present the results of the difference between treatment and control households in the full sample (consisting of both the IHD-compatible as well as the ISKRA households), as presented in Table 2. Columns (4)-(6) presented the differences in the inverse probability weighted characteristics of just the IHD-compatible households (418 in treatment, 233 in control). As shown in column (6), three differences are significant: the share of households with children ($p = 0.090$), the share of households living in apartments ($p = 0.005$) and the share of households living in terraced houses ($p = 0.054$). When correcting for Multiple Hypothesis Testing (see Table B4 in Appendix B, Supplementary Materials), the difference in the type of dwelling households live in is the only characteristic that remains (borderline) significant, with the difference in the share of households living in terraced houses having Bonferroni–Holm and Sidak–Holm corrected p-values of 0.026, and the difference in the share of households living in apartments remaining borderline significant with $p = 0.115$ according to Anderson's sharpened q. All other differences have MHT-corrected p-values of 0.184 or higher.

In principle, estimating the impact of offering an IHD among the subsample of households with IHD-compatible smart meters is straightforward because the allocation to treatment and control was randomized by us researchers — albeit not with a 50-50 chance. However, dropping the ISKRA-metered households from the control group reduces the sample size; to correct for possible imbalances we apply inverse probability weighting (IPW); see Imbens and Wooldridge (2009). For each batch B we calculated the total number of IHD-compatible households (N_C^B, where subscript "C" refers to compatible) as well as the total number of IHD-compatible households that ended up in the treatment group (n_T^B, where subscript "T" refers to treatment). The probabilities of an IHD-compatible household having ended up in the treatment or in the control group in a batch are then equal to, respectively, n_T^B/N_C^B and $(N_C^B - n_T^B)/N_C^B$. We then use the inverse of these two ratios as weights in the calculation of the means of the characteristics of the treatment and control group (to verify whether the two groups are balanced; see Table A.3) and also as weights in the impact analyses (see Table A.4).

Before presenting the IPW-corrected regression results, we first test whether balance on characteristics is still decent after dropping all ISKRA-metered households and subsequently applying IPW. The results are presented in Table A.3. For ease of comparison, columns (1)-(3) of this table replicate the balance test for the full sample (consisting of both all IHD-compatible households and the ISKRA households; cf. Table 2); columns (4)-(6) present the results of the IPW-weighted balance test for just the subsample of IHD-compatible households.

As is clear from columns (4)-(6) of this table, differences in characteristics tend to be small, with the only significant differences being a lower share of households with children below 18 in the treatment group ($p = 0.090$), and the share of households living in apartments being smaller ($p = 0.005$) and the share of households living in terraced houses ($p = 0.054$) being larger in the treatment

Table A.4
Average Treatment Effect estimates of the IHD on the daily amount of gas, electricity and total energy consumed for the subsample of IHD-compatible households, weighted with the inverse of the probability of households ending up in treatment or control.

	Full sample, unweighted			Subsample, IPW-weighted			IPW subsample, excl. Batch # 14		
	(1) Electricity	(2) Gas	(3) Energy	(4) Electricity	(5) Gas	(6) Energy	(7) Electricity	(8) Gas	(9) Energy
Treatment	−134.703**	−0.150**	−1579.503**	−87.845	−0.129**	−1342.422**	−137.310*	−0.113**	−1233.155**
	(61.488)	(0.061)	(615.124)	(69.959)	(0.054)	(541.378)	(79.250)	(0.051)	(505.319)
Control group mean	7660.794	2.665	33,597.507	7528.009	2.656	33,426.824	7458.943	2.631	33,103.103
F-value	12.739	11.154	8.824	200.340	34.709	24.982	147.679	6.952	10.248
# hhs (N)	802	802	802	645	645	645	611	611	611
# Obs. (N × T)	346,321	343,375	341,771	277,431	275,649	274,292	259,987	258,211	256,927
Household fixed effects	Y	Y	Y	Y	Y	Y	Y	Y	Y
Day × region fixed effects	Y	Y	Y	Y	Y	Y	Y	Y	Y
Weather controls	Y	Y	Y	Y	Y	Y	Y	Y	Y

Notes: ATE impact estimates on electricity consumption (in Wh, in columns (1), (4) and (7)), natural gas consumption (in m^3, in columns (2), (5) and (8)) and on total energy consumption (in Wh, in columns (3), (6) and (9)). For ease of comparison, columns (1)-(3) replicate the results of Panel D in Table 3. These regressions contain all control group households (those with but also those without IHD-compatible smart meters), without any weighting. Columns (4)-(6) present the estimates of a regression using only households with IHD-compatible smart meters, where each observation is weighted by the inverse of the probability that the household ended in the treatment or control group. Batch # 14 is the one in Location 6 that was randomized on November 30, 2019. From the 34 IHD-compatible households in that batch only 2 ended up in control, because of the very large share of ISKRA-metered households in that batch; see Table 1. The regression models used in columns (7)-(9) present the IPW-weighed outcomes excluding Batch # 14. Standard errors, presented in parentheses, are clustered either by 2-digit zip code (in those regions with non-updatable ISKRA meters) or at the household level; *, **, and *** indicate significance at, respectively, the 10, 5, and 1 percent critical level. The control group mean is the average amount of energy consumed by the control group households since the start of the intervention, as defined by the median installation date of the IHD in a batch.

group than in the control group. As none of the differences are robust to multiple hypothesis testing (see Table B4 in Appendix B, Supplementary Materials), we conclude that the balance of the IPW-weighted households with IHD-compatible smart meters is decent.

The impact analysis using just the sample of IHD-compatible households and applying IPW is presented in Table A.4. For ease of comparison, columns (1)–(3) of this table present the treatment effect estimates for electricity, gas and total energy consumption as presented in Panel D of Table 3. The revised impact estimates using IPW for just the subsample of IHD-compatible households are presented in columns (4)-(6) of Table A.4. Using inverse probability weighting the impact estimate for gas is slightly smaller than without weighting (compare columns (5) and (2)) but the one for electricity is markedly lower and no longer significant at conventional significance levels (compare columns (1) and (4)). However, the probabilities of ending up in the treatment group differ quite substantially between batches. We took a closer look at one batch (number 14, in Location 6; i.e., the batch of November 30, 2019, see Table 1). This batch consisted of 29 ISKRA households and 34 IHD-compatible households, and of the latter 32 ended up in treatment and 2 in control. The overall share of treated households in the batch is thus about 50 percent (i.e., 32/63), but the weight assigned to each of the two IHD-compatible households in the control group is 16 times larger than the weight assigned to each of the 32 IHD-compatible households in the treatment group. Although assignment to treatment was random within the batch, the number of IHD-compatible households assigned to control (i.e., two) is not large enough to assume that they are a representative subsample of the set of 34 IHD-compatible households from which they were drawn.

We therefore reran the analyses using inverse probability weighting without Batch #14, and the results are presented in columns (7)-(9) of Table A.4. As shown in column (8), the impact estimate for gas is again smaller than in the previous two regressions (compare column (8) to columns (2) and (5)) but remains significant at the 5% level, while the impact estimate for electricity is again very close to (and even weakly larger than) the estimate using the overall sample; compare columns (7) and (1). The decrease in significance of the estimate of the impact on electricity ($p = 0.085$ in column (7) rather than $p = 0.030$ in column (1)) is thus due to the decrease in sample size (from 802 to 611) – not because of IPW. We thus find that our results are by and large robust to omitting the ISKRA households and subsequently using inverse probability weighting on the IHD-compatible households.

A.2.3. Testing for possible biases resulting from treatment imbalances

Table 2 shows that the treatment and control groups are well-balanced. The means of just one of the twenty baseline characteristics differ significantly between treatment and control (the share of elderly households), and the statistical significance of this difference is not robust to multiple hypotheses testing (see Table B3 in Appendix B, Supplementary Materials); also the normalized difference in the share of elderly households is less than 0.25 standard deviations (see column (5) in Table 2). Still, we test whether explicitly controlling for this difference in our regression analyses affects the size of the treatment estimates. The results of the Random Effects model are presented in Table A.5. In columns (1), (3) and (5) we present the results when controlling for just the unbalanced variable (the share of elderly households, i.e. whether one or more members of the household is over 65), while columns (2), (4) and (6) present the results when controlling for all covariates. We lose up to two observations because of missing information. The treatment estimates are very close to the ones using fixed effects; compare the results of Table A.5 to those of Panel D of Table 3.

Table A.5
Random-effects Average Treatment Estimates using either the unbalanced variable (as identified in Table 2, whether one or more members of the household are over 65) or all household characteristics as controls.

	Electricity		Gas		Total energy	
	(1)	(2)	(3)	(4)	(5)	(6)
Treatment	−136.1**	−134.9**	−0.151**	−0.152**	−1589.6**	−1598.2***
	(61.45)	(61.22)	(0.0616)	(0.0617)	(620.7)	(620.4)
Elderly hh	−652.0***	−410.1*	0.114*	0.175***	466.9	1307.7*
	(251.1)	(215.4)	(0.0661)	(0.0671)	(833.8)	(760.4)
1 person hh		−2099.5***		−0.328***		−5270.2***
		(259.1)		(0.0667)		(800.6)
> 2 hh members		1512.0***		0.200**		3460.1***
		(525.4)		(0.0868)		(898.7)
Household with children below 18		1261.4***		0.319***		4390.6***
		(228.0)		(0.102)		(1119.8)
Just primary education		1561.1**		0.183		3352.6*
		(618.5)		(0.164)		(2004.0)
Higher tertiary education		−383.1		0.136		937.9
		(257.5)		(0.0846)		(1019.8)
Dwelling owned by the household		859.9***		0.0143		824.0
		(281.3)		(0.0548)		(754.7)
Apartment		−646.2**		−0.365**		−4279.0***
		(277.1)		(0.144)		(1419.6)
Corner house		586.5*		0.324***		3733.4***
		(343.6)		(0.0878)		(1081.6)
Semi-detached house		806.4**		0.470***		5405.0***
		(348.5)		(0.124)		(1218.1)
Detached house		3269.1***		1.692***		19793.9***
		(387.8)		(0.172)		(1612.7)
Other (farmhouse, houseboat)		5500.7***		2.384***		28839.2***
		(742.5)		(0.366)		(4236.7)
House built after 1974		−136.6		−0.331***		−3334.9***
		(213.6)		(0.0929)		(918.8)
Owned photovoltaic cells at time of randomization		208.5		−0.124		−993.9
		(220.8)		(0.0851)		(967.0)
Control group mean	7660.8	7665.6	2.648	2.668	33 416.2	33 635.5
Number of hhs (N)	801	800	801	800	801	800
Observations (N × T)	346 076	345 746	345 232	342 797	343 628	341 193
Household fixed effects	Y	Y	Y	Y	Y	Y
Day × region effects	Y	Y	Y	Y	Y	Y
Weather controls	Y	Y	Y	Y	Y	Y

Notes: ATE estimates of the IHD's impact on electricity consumption (in Wh, in columns (1)–(2)), gas consumption (in m^3, in columns (3)–(4)) and on the total amount of energy consumed by the household (in Wh, in columns (5)-(6)) using random effects controlling for whether one or more members of the household are over 65 (the one characteristic we failed to have balance on, see Table 2 – in the uneven-numbered columns), or controlling for all baseline characteristics we have information on (in the odd-numbered columns). The total amount of energy consumed is calculated using the Wh content of natural gas in the Netherlands, where the energy content of 1 m^3 of natural gas equals 9,769 Wh. Standard errors, presented in parentheses, are clustered at either the 2-digit zipcode level (in those regions with non-updatable ISKRA meters) or the household level; *, **, and *** indicate significance at, respectively, the 10, 5, and 1 percent critical level. The control group mean is the average amount of energy consumed by the control group households since the start of the intervention, as defined by the median installation date of the IHD in a batch.

A.2.4. Robustness of the results to including outliers

As stated in Section 2.2, eight households in our sample had annual electricity consumption levels of 9000 kWh or more, which is well above the 8150 kWh consumed by the 95th percentile of the group of households using most energy in the Netherlands (those consisting of five or more household members living in large detached houses).[14] With four outliers in treatment and four in control, the allocation of these outliers to treatment and control was balanced; see the numbers in superscripts in the various cells of Table 1. During the baseline period, the average daily amount of electricity consumed by these households was 3.5 as high as that of the rest of the sample, and also the daily amount of gas consumed was 1.5 times as high as that of the rest of the sample. These amounts of electricity consumption are unlikely to be the result of measurement error (as households' energy bills are derived using the metered amounts of gas and electricity consumed, and hence all meters have been carefully calibrated).

In our main analysis we simply dropped these eight households, but this can be considered a crude way of dealing with the issue at hand. Alternatively we can winsorize the data. Table A.6 presents the results of running the analysis using the full sample (i.e., including the eight outliers), as well as those when subsequently winsorizing gas, electricity and total energy consumption, per day, at the 0.5% and 1% level. In this table we applied one-sided winsorization (just at the top, not at the bottom) because very low levels of consumption are unlikely to be outliers in a non-negligible number of days in the year (think of vacation periods);

[14] See https://www.cbs.nl/nl-nl/cijfers/detail/83882ned; website consulted on 10-7-2020.

Table A.6
Average Treatment Estimates of the impact of the IHD on gas, electricity and total energy consumption including outliers — full sample, as well as subsequently winsorized, at the top, at the 0.5% and 1% level.

	No winsorization			Top 0.5% Winsorized			Top 1.0% Winsorized		
	(1) Electricity	(2) Gas	(3) Total energy	(4) Electricity	(5) Gas	(6) Total energy	(7) Electricity	(8) Gas	(9) Total energy
Treatment	−181.706**	−0.112	−1262.181*	−174.235***	−0.117*	−1325.868**	−167.974***	−0.116*	−1319.962**
	(75.557)	(0.075)	(709.592)	(63.980)	(0.070)	(667.433)	(58.236)	(0.067)	(639.223)
Control group mean	7803.512	2.669	33,779.754	7734.959	2.648	33,570.883	7687.287	2.628	33,372.698
F-value	65.356	10.283	7.877	88.484	14.834	7.865	66.639	25.631	8.060
Number of hhs (N)	810	810	810	810	810	810	810	810	810
Observations (N × T)	349,652	346,712	345,098	349,652	346,712	345,098	349,652	346,712	345,098
Household fixed effects	Y	Y	Y	Y	Y	Y	Y	Y	Y
Day × region fixed effects	Y	Y	Y	Y	Y	Y	Y	Y	Y
Weather controls	Y	Y	Y	Y	Y	Y	Y	Y	Y

Notes: Results of the ATE regressions using all households (i.e., including those that we identified as outliers: those with an estimated annual electricity consumption of 9,000 kWh or more, which is above the 95th percentile of the electricity consumption of the subpopulation in the Netherlands with the highest electricity consumption — households of 5 or more members living in detached homes). Columns (1)–(3) present the estimates for electricity, gas and total energy consumption without winsorization; those in columns (4)–(6) and (7)–(9) present those using 0.5% and 1% winsorization, one-sided at the top. Standard errors, presented in parentheses, are clustered either by 2-digit zip code (in those regions with non-updatable ISKRA meters) or at the household level; *, **, and *** indicate significance at, respectively, the 10, 5, and 1 percent critical level. The control group mean is the average amount of energy consumed by the control group households since the start of the intervention, as defined by the median installation date of the IHD in a batch. Results are robust to applying symmetric winsorization; see Table B5 in Appendix B (Supplementary Materials).

see Table O5 in the Online Appendix for the results applying two-sided winsorization. From Table A.6 we infer that with the eight outliers in the analysis the estimated ATE impact on electricity is larger than when excluding them, independent of whether the outcome variables are winsorized, or by how much; compare columns (1), (4) and (7) in Table A.6 to those of column (1) in Panel D of Table 3. Including the eight outliers results in estimated gas savings that are lower than when excluding them (compare columns (2), (5) and (8) in Table A.6 to those of column (2) in Panel D of Table 3) but the difference remains significant when winsorizing, one-sided at the top, at the 0.5% and 1% level (without winsorizing the savings fail to be significant at p-value of 0.138). The same conclusions apply to the case of two-sided winsorization; see Table O5 in the Online Appendix.

So our results are by and large robust to winsorizing the data. Note, however, that the eight households that we labeled outliers occur in just 5 out of the 25 randomization batches; see Table 1. This does not represent a bias as Locations 5 and 6 are also the ones with the largest number of participant households (with respectively 187 and 298 households), and July 10, 2019 was also the date on which almost half of the households started. The uneven distribution of outliers between the randomization batches does mean, however, that winsorizing the data is too strict – or even unnecessary – on many days (up to three months before July 10, 2019 and onwards from the twelfth month thereafter) and in most locations (in all locations other than 5 and 6), and not strict enough on the days around July 10, 2019 and in Locations 5 and 6. The impact estimates presented in Table 3 are thus likely to be more accurate than those presented in Table A.6.

A.2.5. Robustness to the definition of the start of the treatment period

The definition of the start of the intervention period is straightforward for those treatment households that accepted the IHD; for these households the intervention starts at the date of installation of their IHD. The definition of the start of the intervention period is less clear-cut for those households in the treatment group that ended up not accepting the IHD. The definition of the start of the intervention period for these"non-compliers" is important as it is likely to affect the size treatment estimates, and the definition of the start of the intervention period is also important for the calculation of the counterfactual amount of energy consumed (and hence for the percentage changes in the amount of energy consumed) – i.e., the average amount consumed by the control group households in the intervention period.

The results presented in Tables 3 and 4 were obtained using the IHD's median installation date in a batch as the definition of the start of the intervention period for those treatment households that did not end up with an IHD. We also tested whether our results are robust to an alternative definition of the start of the intervention period: the date at which the *first* treatment household in a batch received their IHD. The results of this analysis are presented in Table A.7. Comparing these results to those in Panel D of Tables 3 and 4, the treatment estimates are qualitatively identical and quantitatively very similar.

A.2.6. Relevance of correcting for the production of electricity using solar panels

The electricity consumption's impact estimates presented in Tables 3 and 4 were obtained upon correcting the net amount of electricity purchased by an estimate of the amount of electricity produced, for every day, by households owning solar panels. We reran the same regressions without this correction; see Table A.8. Dropping the correction for self-consumption means that the unit of measurement changes from the (estimated) consumption of electricity to the net amount purchased. The amount purchased is equal to the amount consumed for those households without solar panels, but not for the approximately 40% of the households in the sample with a solar panel system; see Table 2. Without the correction the impact estimates are in terms of reductions in the net amount purchased which is, for each household, weakly smaller than the reductions in the total amount consumed.

Table A.7
Average Treatment Effect (ATE) and the Average Treatment-on-the-Treated (ATT) effect estimates of the impact of the IHD on the daily amount of gas (in m^3) and electricity (in Wh) consumed as well as on the total amount of energy consumed (in Wh), using the date of installation of the first IHD installed in a batch as the starting date of the intervention period.

	Electricity		Gas		Total energy	
	(1)	(2)	(3)	(4)	(5)	(6)
Treatment (ATE) or instrumented installation (ATT)	−143.990** (63.960)	−170.681** (77.580)	−0.147*** (0.049)	−0.175*** (0.057)	−1562.677*** (493.967)	−1853.731*** (579.008)
Estimate	ATE	ATT	ATE	ATT	ATE	ATT
Control group mean	7745.881	7745.881	2.547	2.547	32,543.015	32,543.015
F-value	12.365	12.189	11.458	11.451	8.795	8.887
Number of hhs (N)	802	802	802	802	802	802
Observations (N × T)	346,321	346,321	343,375	343,375	341,771	341,771
Household fixed effects	Y	Y	Y	Y	Y	Y
Day × region fixed effects	Y	Y	Y	Y	Y	Y
Weather controls	Y	Y	Y	Y	Y	Y
Kleibergen–Paap rk LM statistic		8.731		8.758		8.755

Notes: ATE and ATT estimates, in respectively columns (1), (3) and (5) and columns (2), (4) and (6), on electricity consumption (in Wh, in columns (1) and (2)), gas consumption (in m^3, in columns (3) and (4)) and on the total amount of energy consumed by the household (in Wh, in columns (5) and (6)). The total amount of energy consumed is calculated using the Wh content of natural gas in the Netherlands, where 1 m^3 of natural gas equals 9,769 Wh. The ATT estimates (or Local Average Treatment Effects, LATE) correct for imperfect take-up of the IHD among the treatment group households. The control group mean is the average amount of energy consumed by the control group households since the start of the intervention, here defined by the date at which the first IHD was installed in a batch. The Kleibergen–Paap statistics in columns (2), (4) and (6) are significantly different from zero ($p < 0.01$), indicating that our estimates do not suffer from weak identification bias. Standard errors, presented in parentheses, are clustered either at the 2-digit zip code level (in those regions with non-updatable ISKRA meters) or at the household level. *, **, and *** indicate significance at, respectively, the 10, 5, and 1 percent critical level.

Table A.8
Estimating the impact of the IHD on electricity consumption without correcting for self-consumption of own electricity production.

	Electricity	
	(1)	(2)
Treatment or instrumented installation	−115.742+ (74.150)	−143.051+ (90.985)
Estimate	ATE	ATT
Control group mean	7091.133	7091.133
F-value	1.507	1.517
Number of hhs (N)	802	802
Observations (N × T)	346,321	346,321
Household fixed effects	Y	Y
Day × region fixed effects	Y	Y
Weather controls	Y	Y
Kleibergen–Paap rk LM statistic		8.793

Notes: ATE and ATT impact estimates, in respectively columns (1) and (2), of the impact of the IHD on the net amount of electricity purchased – that is, without correcting for the amount households consume of the electricity produced by their solar panel system. The ATT estimate (or Local Average Treatment Effect, LATE) corrects for imperfect take-up of the IHD among the treatment group households. The Kleibergen–Paap statistic in column (2) is significantly different from zero ($p < 0.01$), indicating that our estimates do not suffer from weak identification bias. Standard errors, presented in parentheses, are clustered either at the 2-digit zip code level (in those regions with non-updatable ISKRA meters) or at the household level; +, *, **, and *** indicate significance at, respectively, the 12.5, 10, 5, and 1 percent critical level. The control group mean is the average amount of energy consumed by the control group households since the start of the intervention, as defined by the median installation date of the IHD in a batch.

Table A.8 shows that without correcting for self-consumption of the amount of electricity produced the treatment estimates are, in absolute terms, smaller than those presented in Table 3, and they also fail to be significant, albeit just so, at the 10% level; the *p*-values are 0.120 and 0.117 for the ATE and ATT estimates, respectively. But because the denominators are smaller too (because self-consumption of the electricity one produced is ignored), the percentage changes are by and large unchanged as they are 1.63% and 2.02% for the ATE and ATT, whereas they were 1.76% and 2.17% in Tables 3 and 4, respectively; see Section 3.1. Based on these results we conclude that having taken the effort to correct for self-consumption was worth the effort; in percentage terms the savings are very similar, but the correction for the amount of electricity produced (and hence for self-consumption) yields more precise estimates (and hence smaller standard errors).

A.2.7. Testing for the presence of heterogeneous treatment effects

The typical finding is that the treatment effect of information provision tends to be larger among those households with higher baseline levels of energy consumption, outdoor temperatures, and home ownership status; see for example Abrahamse et al. (2007),

Table A.9
Heterogeneous Average Treatment Effects for above- versus below-median intensity of baseline energy consumption, home ownership status, and heating and/or cooling season.

	Basel. en. cons.	Home ownership	Outdoor temp.
	(1)	(2)	(3)
Panel A: Electricity (Wh)			
Treatment	20.517	−165.434*	−129.167*
	(70.502)	(95.514)	(70.075)
Treatment ×	−188.122	85.699	−10.512
characteristic	(155.816)	(293.854)	(117.295)
Control group mean	7660.794	7660.794	7660.794
F-value	13.920	9.196	11.866
Panel B: Gas (m^3)			
Treatment	−0.164**	−0.204**	−0.112*
	(0.077)	(0.093)	(0.061)
Treatment ×	−0.027	0.116	−0.058
characteristic	(0.118)	(0.185)	(0.109)
Control group mean	2.665	2.665	2.665
F-value	9.151	15.642	12.776
Panel C: Total energy (Wh)			
Treatment	−1649.532*	−2152.215**	−1107.684**
	(879.289)	(942.454)	(458.259)
Treatment ×	−175.048	1346.365	−598.324
characteristic	(1300.076)	(1915.024)	(556.880)
Control group mean	33,597.507	33,597.507	33,597.507
F-value	7.618	14.057	8.984
Number of hhs (N)	802	801	802
Observations (N × T)	346,321	345,990	346,321
Household fixed effects	Y	Y	Y
Day × region fixed effects	Y	Y	Y
Dummy post-install. ×char.	Y	Y	Y
Weather controls	Y	Y	Y

Notes: Heterogeneous treatment effect estimates (ATE) obtained controlling for the full set of levels and interactions terms of treatment status, a dummy indicating the post-intervention period, and the potential heterogeneity characteristic; see equation (A.1). Dependent variables are the amount of electricity, gas and total energy consumed per household per day presented in, respectively, Panels A, B and C. To control for seasonality, the cooling (heating) season consists of the warmest (coolest) 25% of the days of the period during which a household's energy consumption was measured. The control group mean is the average amount of energy consumed by the control group households since the start of the intervention, as defined by the median installation date of the IHD in a batch. Standard errors, presented in parentheses, are clustered either by 2-digit zip code (in those regions with non-updatable ISKRA meters) or at the household level. *, **, and *** indicate significance at, respectively, the 10, 5, and 1 percent critical level.

Costa and Kahn (2013), Asensio and Delmas (2015), Byrne et al. (2018), Tiefenbeck et al. (2018), Andor et al. (2020), Gerster et al. (2021). We follow this literature and test for the presence of heterogeneous treatment effects (ATE) using the following model:

$$y_{it} = \beta_{ATE} T_{it} + \beta_{ATEh} T_{it} I(char)_{it} + \zeta_h I(char)_{it} POST_t + \delta X_{rt} + \alpha_i + \gamma_{rt} + \epsilon_{it}. \quad \text{(A.1)}$$

In this model, $I(char)_{it}$, with $char = \{BaselineEnergyConsumption, HomeOwner, OutdoorTemperature\}$, is a dummy variable that is equal to one if household i is an above-median intensive consumer of the energy source under consideration ($I(BaselineEnergyConsumption)_{it} = 1$ for all t) or if household i owns their dwelling rather than rent it ($I(HomeOwner)_{it} = 1$ for all t) or whether the household's gas (electricity) consumption is observed in the coldest (hottest) quartile of the year ($I(OutdoorTemperature)_{it} = 1$ for all t in either the coldest or the hottest season) – and zero otherwise. The coding of the season effects differ between gas and electricity as gas is typically used for heating (and hence in the coldest quartile of the year) while electricity can be used for cooling (and hence is relevant in the hottest quartile of the year, even though air conditioning was not very comment yet at the time we implemented our study). Note that by including $I(char)_{it} \times POST_t$ we allow energy consumption to differ between each of the two subgroups within $char$ not just in the baseline period (as captured by the household fixed effect), but also in the intervention period.

The results are presented in Table A.9. We find that none of the interaction terms are significant, although there is some weak evidence that treatment effect on electricity consumption is larger for households with above-median electricity consumption at baseline; see the results in column (1) of Panel A of this table.

A.3. Characteristics of the survey respondents and their representativeness of the overall sample

The overall response rate to the survey was 55%. Table A.10 provides insight into how representative these respondent households are of the overall sample of participants. Column (2) of the table presents the characteristics of the households that

Table A.10
Representativeness of the survey respondents with respect to the sample of participant households.

Variable	(1)		(2)		(3)		Absolute	Normalized
	Non-respondent		Respondent		Total		difference	difference
	N	Mean/SE	N	Mean/SE	N	Mean/SE	(1)–(2)	(1)–(2)
1 person hh	359	0.304(0.024)	442	0.301(0.022)	801	0.302(0.016)	0.003	0.006
2 person hh	359	0.563(0.026)	442	0.570(0.024)	801	0.567(0.018)	−0.007	−0.015
> 2 hh members	359	0.134(0.018)	442	0.129(0.016)	801	0.131(0.012)	0.005	0.014
Elderly hh	359	0.212(0.022)	442	0.303(0.022)	801	0.262(0.016)	−0.091***	−0.208
Household with children below 18	360	0.353(0.025)	442	0.330(0.022)	802	0.340(0.017)	0.022	0.047
Just primary education	359	0.039(0.010)	442	0.016(0.006)	801	0.026(0.006)	0.023**	0.145
Secondary or lower tertiary education	359	0.585(0.026)	442	0.484(0.024)	801	0.529(0.018)	0.101***	0.202
Higher tertiary education	359	0.376(0.026)	442	0.500(0.024)	801	0.444(0.018)	−0.124***	−0.249
Apartment	359	0.248(0.023)	442	0.247(0.021)	801	0.247(0.015)	0.001	0.003
Terraced house	359	0.287(0.024)	442	0.242(0.020)	801	0.262(0.016)	0.045	0.102
Corner house	359	0.178(0.020)	442	0.199(0.019)	801	0.190(0.014)	−0.021	−0.053
Semi-detached house	359	0.170(0.020)	442	0.163(0.018)	801	0.166(0.013)	0.007	0.019
Detached house	359	0.086(0.015)	442	0.145(0.017)	801	0.119(0.011)	−0.058**	−0.181
Other (farmhouse, houseboat)	359	0.031(0.009)	442	0.005(0.003)	801	0.016(0.004)	0.026***	0.207
Dwelling owned by the household	359	0.423(0.026)	442	0.455(0.024)	801	0.441(0.018)	−0.031	−0.063
House built after 1974	358	0.366(0.025)	442	0.382(0.023)	800	0.375(0.017)	−0.016	−0.034
Owned photovoltaic cells at time of randomization	360	0.356(0.025)	442	0.452(0.024)	802	0.409(0.017)	−0.097***	−0.197
Baseline average daily gas consumption (m3)	360	1.721(0.108)	442	1.607(0.100)	802	1.658(0.073)	0.115	0.055
Baseline average daily electricity consumption (kWh)	360	7.963(0.234)	442	7.529(0.210)	802	7.724(0.156)	0.434	0.098
Baseline average daily energy expenditures, in Euros	360	2.711(0.096)	442	2.550(0.088)	802	2.622(0.065)	0.161	0.087

Notes: Descriptive statistics of the those households that filled out the survey (in column (2)) and those who did not (in column (1)) as well as the absolute and normalized differences between the two (in columns (4) and (5), respectively); standard errors are presented in parentheses. The table was constructed using the iebaltab command of the STATA's ietoolkit package (DIME Analytics, 2019). *, **, and *** indicate significance at, respectively, the 10, 5, and 1 percent critical level. Some of the normalized differences are close to the threshold of 0.25 standard deviations, and as shown in Table B6 in Appendix B (Supplementary Materials) four (groups of) characteristics are still significant upon correcting for Multiple Hypotheses Testing (MHT): respondents are older, more likely to have higher tertiary (as opposed to secondary and lower tertiary) education, they are more likely to live in a detached house and they are also more likely to have solar panels at baseline.

Table A.11
Characteristics of the survey respondents, and the differences between treatment and control households among them.

Variable	(1)		(2)		(3)		(4)	(5)
	Control group		Treatment group		Total		Difference	Normalized diff.
	N	Mean/SE	N	Mean/SE	N	Mean/SE	(1)–(2)	(1)–(2)
1 person hh	183	0.322	259	0.286	442	0.301	0.037	0.080
		(0.035)		(0.028)		(0.022)		
2 person hh	183	0.541	259	0.591	442	0.570	−0.050	−0.100
		(0.037)		(0.031)		(0.024)		
> 2 hh members	183	0.137	259	0.124	442	0.129	0.013	0.039
		(0.025)		(0.020)		(0.016)		
Elderly hh	183	0.240	259	0.347	442	0.303	−0.107**	−0.233
		(0.032)		(0.030)		(0.022)		
Household with children below 18	183	0.355	259	0.313	442	0.330	0.042	0.090
		(0.035)		(0.029)		(0.022)		
Just primary education	183	0.016	259	0.015	442	0.016	0.001	0.008
		(0.009)		(0.008)		(0.006)		
Secondary or lower tertiary education	183	0.486	259	0.483	442	0.484	0.004	0.007
		(0.037)		(0.031)		(0.024)		
Higher tertiary education	183	0.497	259	0.502	442	0.500	−0.005	−0.009
		(0.037)		(0.031)		(0.024)		
Apartment	183	0.197	259	0.282	442	0.247	−0.085**	−0.197
		(0.029)		(0.028)		(0.021)		
Terraced house	183	0.262	259	0.228	442	0.242	0.034	0.080
		(0.033)		(0.026)		(0.020)		
Corner house	183	0.230	259	0.178	442	0.199	0.052	0.130
		(0.031)		(0.024)		(0.019)		
Semi-detached house	183	0.169	259	0.158	442	0.163	0.011	0.030
		(0.028)		(0.023)		(0.018)		
Detached house	183	0.142	259	0.147	442	0.145	−0.005	−0.013
		(0.026)		(0.022)		(0.017)		
Other (farmhouse, houseboat)	183	0.000	259	0.008	442	0.005	−0.008	−0.115
		(0.000)		(0.005)		(0.003)		
Dwelling owned by the household	183	0.459	259	0.452	442	0.455	0.007	0.015
		(0.037)		(0.031)		(0.024)		
House built after 1974	183	0.350	259	0.405	442	0.382	−0.056	−0.114
		(0.035)		(0.031)		(0.023)		
Owned photovoltaic cells at time of randomization	183	0.459	259	0.448	442	0.452	0.011	0.022
		(0.037)		(0.031)		(0.024)		
Baseline average daily gas consumption (m3)	183	1.679	259	1.556	442	1.607	0.123	0.059
		(0.156)		(0.130)		(0.100)		
Baseline average daily electricity consumption (kWh)	183	7.651	259	7.443	442	7.529	0.208	0.047
		(0.322)		(0.277)		(0.210)		
Baseline average daily energy expenditures, in Euros	183	2.621	259	2.500	442	2.550	0.122	0.066
		(0.135)		(0.116)		(0.088)		

Notes: Overall statistics (means, and standard errors) of the sample of households that filled out the survey, and the results of the balance tests (using absolute and normalized differences) between the treatment and control households among the survey respondents. The table was constructed using the iebaltab command of the STATA's ietoolkit package (DIME Analytics, 2019). *, **, and *** indicate significance at, respectively, the 10, 5, and 1 percent critical level. When correcting for Multiple Hypotheses Testing (MHT) all absolute differences are insignificant (with $p = 0.273$ or higher) according to a series of MHT tests (Anderson sharpened q, Bonferroni–Holm, Sidak–Holm and Westfall–Young (Westfall and Young, 1993); see Table B7 in Appendix B, Supplementary Materials), and also none of the normalized differences are larger than 0.25 standard deviations.

filled out the survey, whereas the characteristics of the non-respondents are summarized in column (1). Households that responded were more likely to be elderly, were higher educated (i.e., more likely to have higher tertiary education, and less likely to have lower levels of education), more likely to live in a detached (and hence typically larger) house and to have had solar panels at baseline. These differences are likely to be relevant, as the normalized differences for these characteristics are typically large (i.e., not much lower than 0.25 standard deviations, the critical threshold typically used in this literature above which subsamples are considered to be unbalanced), and these four characteristics also remain significant even after correcting for multiple hypotheses testing; see Table B6 in Appendix B (Supplementary Materials).

Next, Table A.11 provides insight into the extent to which households of the two experimental groups are represented equally among the subsample of survey respondents. Comparing columns (1) and (2) of this table, we find that the share of elderly households is larger in the treatment than in the control group ($p = 0.016$), and this is also the case for the share of households living in apartments ($p = 0.041$). However, these differences are not robust to multiple hypotheses testing; see Table B7 in Appendix B (Supplementary Materials). For all other household and home characteristics, differences are small and statistically insignificant, and none of the normalized differences are larger than 0.25.

Appendix B. Supplementary materials

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.jeem.2025.103163.

References

Abadie, A., Athey, S., Imbens, G.W., Wooldridge, J.M., 2023. When should you adjust standard errors for clustering? Q. J. Econ. 138 (1), 1–35.

Abadie, A., Imbens, G.W., 2011. Bias-corrected matching estimators for average treatment effects. J. Bus. Econom. Statist. 29 (1), 1–11.

Abdi, H., 2007. Bonferroni and Sidak corrections for multiple comparisonsidak corrections for multiple comparisons. Encycl. Meas. Stat. 3, 103–107.

Abrahamse, W., Steg, L., Vlek, C., Rothengatter, T., 2005. A review of intervention studies aimed at household energy conservation. J. Environ. Psychol. 25 (3), 273–291.

Abrahamse, W., Steg, L., Vlek, C., Rothengatter, T., 2007. The effect of tailored information, goal setting, and tailored feedback on household energy use, energy-related behaviors, and behavioral antecedents. J. Environ. Psychol. 27 (4), 265–276.

Agarwal, R., Garg, M., Tejaswini, D., Garg, V., Srivastava, P., Mathur, J., Gupta, R., 2023. A review of residential energy feedback studies. Energy Build. 290, 113071.

Allcott, H., 2011a. Consumers' perceptions and misperceptions of energy costs. Am. Econ. Rev.: Pap. Proc. 101 (3), 98–104.

Allcott, H., 2011b. Social norms and energy conservation. J. Publ. Econ. 95 (9-10), 1082–1095.

Allcott, H., Greenstone, M., 2012. Is there an energy efficiency gap? J. Econ. Perspect. 26 (1), 3–28.

Allcott, H., Mullainathan, S., Taubinsky, D., 2014. Energy policy with externalities and internalities. J. Publ. Econ. 112, 72–88.

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

Anderson, M.L., 2008. Multiple inference and gender differences in the effects of early intervention: A reevaluation of the abecedaian, perry preschool, and early training projectsbecedaian, perry preschool, and early training projects. J. Amer. Statist. Assoc. 103, 1481–1495.

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

Andor, M.A., Gerster, A., Peters, J., 2020. Information Campaigns for Residential Energy Conservation. RWI, Working Paper 871.

Andor, M., Gerster, A., Peters, J., Schmidt, C., 2020. Social norms and energy conservation beyond the US. J. Environ. Econ. Manag. 103, 102351.

Angrist, J.D., Bettinger, E., Bloom, E., King, E., Kremer, M., 2002. Vouchers for private schooling in Colombia: Evidence from a randomized natural experimentolombia: Evidence from a randomized natural experiment. Am. Econ. Rev. 92 (5), 1535–1558.

Angrist, J.D., Imbens, G.W., Rubin, D.B., 1996. Identification of causal effects using instrumental variables. J. Amer. Statist. Assoc. 91 (434), 444–455.

Asensio, O.I., Delmas, M.A., 2015. Nonprice incentives and energy conservation. Proc. Natl. Acad. Sci. 112 (6), E510–E515.

Asensio, O.I., Delmas, M.A., 2016. The dynamics of behavior change: Evidence from energy conservation. J. Econ. Behav. Organ. 126, 196–212.

Attari, S.Z., DeKay, M.L., Davidson, C.I., Bruine de Bruin, W., 2010. Public perceptions of energy consumption and savings. Proc. Natl. Acad. Sci. 107 (37), 16054–16059.

Aydin, E., Brounen, D., Kok, N., 2018. Information provision and energy consumption: Evidence from a field experiment. Energy Econ. 71, 403–410.

Aydin, E., Kok, N., Brounen, D., 2017. Energy efficiency and household behavior: The rebound effect in the residential sectorhe rebound effect in the residential sector. Rand J. Econ. 48 (3), 749–782.

Bandura, A., 1977. Self-efficacy: Toward a unifying theory of behavioral change. Psychol Rev 84, 191–215.

Bandura, A., 1988. Self regulation of motivation and action through goal systems. In: Hamilton, V., Bower, G., Frijda, N. (Eds.), Cognitive Perspectives on Emotion and Motivation. Kluwer Academic Publishers, Dordrecht.

Barrios, T., Diamond, R., Imbens, G.W., Kolesár, M., 2012. Clustering, spatial correlations, and randomization inference. J. Amer. Statist. Assoc. 107 (498), 578–591.

Benhabib, J., Bisin, A., 2005. Modeling internal commitment mechanisms and self-control: A neuroeconomics approach to consumption-saving decisions. Games Econom. Behav. 52 (2), 460–492.

Bloombaum, M., 1983. The Hawthorne experiments: A critique and reanalysis of the first statistical interpretation by Franke and Kaulawthorne experiments: A critique and reanalysis of the first statistical interpretation by Franke and Kaul. Sociol. Perspect. 26, 71–88.

Brounen, D., Kok, N., Quigley, J.M., 2013. Energy literacy, awareness, and conservation behavior of residential households. Energy Econ. 38, 42–50.

Brülisauer, M., Goette, L., Jiang, Z., Schmitz, J., Schubert, R., 2020. Appliance-specific feedback and social comparisons: Evidence from a field experiment on energy conservation. Energy Policy 145, 111742.

Burlig, F., Preonas, L., Woerman, M., 2017. PCPANEL: Stata module to perform power calculations for randomized experiments with panel data, allowing for arbitrary serial correlation. In: Statistical Software Components. Boston College Department of Economics.

Burlig, F., Preonas, L., Woerman, M., 2020. Panel data and experimental design. J. Dev. Econ. 144, 102458.

Byrne, D., Nauze, A., Martin, L., 2018. Tell me something I don't already know: Informedness and the impact of information programs don't already know: Informedness and the impact of information programs. Rev. Econ. Stat. 100, 510–527.

Carlsson, F., Gravert, C., Johansson-Stenman, O., Kurz, V., 2021. The use of green nudges as an environmental policy instrument. Rev. Environ. Econ. Policy 15 (2), 216–237.

CER, 2011. Gas Customer Behaviour Trial Findings Report. Commission for Energy Regulation, Ireland, (Information Paper No. CER11180a).
Costa, D.L., Kahn, M.E., 2013. Energy conservation "nudges" and environmentalist ideology: Evidence from a randomized residential electricity field experiment. J. Eur. Econ. Assoc. 11 (3), 680–702.
Darby, S., 2006. The effectiveness of feedback on energy consumption. In: A Review for DEFRA of the Literature on Metering, Billing and Direct Displays Review for DEFRA of the Literature on Metering, Billing and Direct Displays. Environmental Change Institute.
Degen, K., Efferson, C., Frei, F., Goette, L., Lalive, R., 2013. Smart metering, beratung oder sozialer vergleich: Was beeinflusst den elektrizitätsverbrauch? In: Final Report To the Swiss Federal Office of Energy. Switzerland: Ittigen.
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.
Deryugina, T., MacKay, A., Reif, J., 2020. The long-run dynamics of electricity demand: Evidence from municipal aggregation. Am. Econ. J.: Appl. Econ. 12, 86–114.
DIME Analytics, 2019. IETOOLKIT: Stata module providing commands specially developed for impact evaluations.: Stata module providing commands specially developed for impact evaluations.
Ehrhardt-Martinez, K., Donnelly, K.A., Laitner, J.A., 2010. Advanced Metering Initiatives and Residential Feedback Programs: A Meta-Review for Household Electricity-Saving Opportunities Meta-Review for Household Electricity-Saving Opportunities Report E105. ACEEE, Washington D.C.
European Union, 2019. Directive (EU) 2019/944 of the European parliament and of the council of 5 june 2019 on common rules for the internal market for electricity and amending directive 2012/27/EU. Off. J. Eur. Union.
Faruqui, A., Arritt, K., Sergici, S., 2017. The impact of advanced metering infrastructure on energy conservation: A case study of two utilities. Electr. J. 30 (3), 56–63.
Faruqui, A., Sergici, S., Sharif, A., 2010. The impact of informational feedback on energy consumption: A survey of the experimental evidence survey of the experimental evidence. Energy 35 (4), 1598–1608.
Fischer, C., 2008. Feedback on household electricity consumption: A tool for saving energy? Energy Effic. 1 (1), 79–104.
Fisher, R., 1925. The Design of Experiments. Oliver and Boyd, London.
Frederiks, E., Stenner, K., Hobman, E., 2015. The socio-demographic and psychological predictors of residential energy consumption: A comprehensive review comprehensive review. Energies 8, 573–609.
Gerster, A., Andor, M.A., Goette, L., 2021. Disaggregate Consumption Feedback and Energy Conservation. Sonderforschungsbereich 823, TU Dortmund, Working Paper.
Gillingham, K., Palmer, K., 2014. Bridging the energy efficiency gap: Policy insights from economic theory and empirical evidence. Rev. Environ. Econ. Policy 8 (1), 18–38.
Hargreaves, T., Nye, M., Burgess, J., 2010. Making energy visible: A qualitative field study of how householders interact with feedback from smart energy monitors. Energy Policy 38 (10), 6111–6119.
Hargreaves, T., Nye, M., Burgess, J., 2013. Keeping energy visible? Exploring how householders interact with feedback from smart energy monitors in the longer term. Energy Policy 52, 126–134.
Harold, J., Lyons, S., Cullinan, J., 2015. The determinants of residential gas demand in Ireland. Energy Econ. 51, 475–483.
Harrison, G.G., List, J.A., 2004. Field experiments. J. Econ. Lit. 42 (4), 1009–1055.
Henry, M.L., Ferraro, P.J., Kontoleon, A., 2019. The behavioural effect of electronic home energy reports: Evidence from a randomised field trial in the United States. Energy Policy 132, 1256–1261.
Hess, S., 2017. Randomization inference with stata: A guide and softwaretata: A guide and software. Stata J. 17 (3), 630–651.
Holtom, B., Baruch, Y., Aguinis, H., Ballinger, G.A., 2022. Survey response rates: Trends and a validity assessment framework. Hum. Relat. 75 (8), 1560–1584.
Imbens, G.W., Angrist, J.D., 1994. Identification and estimation of local average treatment effectsocal average treatment effects. Econometrica 62 (2), 467–475.
Imbens, G.W., Rubin, B.D., 2015. Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction. Cambridge University Press, Cambridge.
Imbens, G.W., Wooldridge, J.M., 2009. Recent developments in the econometrics of program evaluation. J. Econ. Lit. 47, 5–86.
Ito, K., 2014. Do consumers respond to marginal or average price? Evidence from nonlinear electricity pricingvidence from nonlinear electricity pricing. Am. Econ. Rev. 104 (2), 537–563.
Jessoe, K., Rapson, D., 2014. Knowledge is (less) power: Experimental evidence from residential energy use. Am. Econ. Rev. 104 (4), 1417–1438.
Jones, D., Molitor, D., Reif, J., 2019. What do workplace wellness programs do? Evidence from the illinois workplace wellness study. Q. J. Econ. 134 (4), 1747–1791.
Karlin, B., Zinger, J.F., Ford, R., 2015. The effects of feedback on energy conservation: A meta-analysis. Psychol. Bull. 141 (6), 1205–1227.
Khanna, T., Baiocchi, G., Callaghan, M., Creutzig, F., Guias, H., Haddaway, N., Minx, J., 2021. A multi-country meta-analysis on the role of behavioural change in reducing energy consumption and CO2 emissions in residential buildings emissions in residential buildings. Nat. Energy 6 (9), 925–932.
Kling, J.R., Liebman, J.B., Katz, L.F., 2007. Experimental analysis of neighborhood effects. Econometrica 75, 83–119.
Lauren, N., Fielding, K.S., Smith, L., Louis, W.R., 2016. You did so you can and you will: Self-efficacy as a mediator of spillover from easy to more difficult pro-environmental behaviour. J. Environ. Psychol. 48, 191–199.
Levitt, S.D., List, J.A., 2007. What do laboratory experiments measuring social preferences reveal about the real world. J. Econ. Perspect. 21, 153–174.
List, J.A., 2024. Optimally generate policy-based evidence before scaling. Nature 626, 491–499.
Luthander, R., Widén, J., Nilsson, D., Palm, J., 2015. Photovoltaic self-consumption in buildings: A review. Appl. Energy 142, 80–94.
Lynham, J., Nitta, K., Saijo, T., Tarui, N., 2016. Why does real-time information reduce energy consumption? Energy Econ. 54, 173–181.
McCloskey, D.N., Ziliak, S.T., 1996. The standard error of regressions. J. Econ. Lit. 34 (1), 97–114.
McKerracher, C., Torriti, J., 2013. Energy consumption feedback in perspective: Integrating Australian data to meta-analyses on in-home displaysustralian data to meta-analyses on in-home displays. Energy Effic. 6 (2), 387–405.
OECD, 2023. How Green Is Household Behaviour?. Organisation for Economic Co-operation and Development, Paris.
Pellerano, J.A., Price, M.K., Puller, S.L., Sánchez, G.E., 2017. Do extrinsic incentives undermine social norms? Evidence from a field experiment in energy conservation. Environ. Resour. Econ. 67 (3), 413–428.
Pischke, S., 2007. Lecture Notes on Measurement Error. London School of Economics, London.
Podgornik, A., Sucic, B., Blazic, B., 2016. Effects of customized consumption feedback on energy efficient behaviour in low-income households. J. Clean. Prod. 130, 25–34, Special Volume: SDEWES 2014 - Sustainable Development of Energy, Water and Environment Systems.
Rindfuss, R.R., Choe, M.K., Tsuya, N.O., Bumpass, L.L., Tamaki, E., 2015. Do low survey response rates bias results? Evidence from Japan. Demogr. Res. 32, 797–828.
RVO, 2018. Marktbarometer Aanbieding Slimme Meters. Netherlands Enterprise Agency (RVO), The Hague.
Saretsky, G., 1972. The OEO P.C. experiment and the John Henry effect experiment and the John Henry effect. Phi Delta Kappan 53, 579–581.
Schleich, J., Faure, C., Klobasa, M., 2017. Persistence of the effects of providing feedback alongside smart metering devices on household electricity demand. Energy Policy 107, 225–233.
Schultz, P., Estrada, M., Schmitt, J., Sokoloski, R., Silva-Send, N., 2015. Using in-home displays to provide smart meter feedback about household electricity consumption: A randomized control trial comparing kilowatts, cost, and social norms randomized control trial comparing kilowatts, cost, and social norms. Energy 90, 351–358.

Schultz, P., Nolan, J., Cialdini, R., Goldstein, N., 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 awarenessawthorne effect and energy awareness. Proc. Natl. Acad. Sci. 110 (38), 15242–15246.
Sexton, S., 2015. Automatic bill payment and salience effects: Evidence from electricity consumption. Rev. Econ. Stat. 97 (2), 229–241.
Shenhav, A., Musslick, S., Lieder, F., Kool, W., Griffiths, T., Cohen, J., Botvinick, M., 2017. Toward a rational and mechanistic account of mental effort. Annu. Rev. Neurosci. 40, 99–124.
Sovacool, B.K., Hook, A., Sareen, S., Geels, F.W., 2021. Global sustainability innovation and governance dynamics of national smart electricity meter transitions. Glob. Environ. Chang. 68, 102272.
Sun, L., Abraham, S., 2021. Estimating dynamic treatment effects in event studies with heterogeneous treatment effects. J. Econometrics 225, 175–199.
Tiefenbeck, V., 2016. On the magnitude and persistence of the Hawthorne effect: Evidence from four field studiesawthorne effect: Evidence from four field studies. In: Proceedings of the 4th European Conference on Behaviour and Energy Efficiency.
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. Manag. Sci. 64 (3), 1458–1476.
Tiefenbeck, V., Staake, T., Roth, K., Sachs, O., 2013. For better or for worse? Empirical evidence of moral licensing in a behavioral energy conservation campaignmpirical evidence of moral licensing in a behavioral energy conservation campaign. Energy Policy 57, 160–171.
Tiefenbeck, V., Wörner, A., Schöb, S., Fleisch, E., Staake, T., 2019. Real-time feedback reduces energy consumption among the broader public without financial incentives. Nat. Energy 4 (10), 831–832.
Trachtman, H., 2024. Does promoting one behavior distract from others? Evidence from a field experiment. AEJ Appl. Econ. 16 (2), 249–277.
Vitiello, S., Andreadou, N., Ardelean, M., Fulli, G., 2022. Smart metering roll-out in Europe: Where do we stand? Cost benefit analyses in the clean energy package and research trends in the green dealurope: Where do we stand? Cost benefit analyses in the clean energy package and research trends in the green deal. Energies 15 (2340).
Westfall, P., Young, S., 1993. Resampling-Based Multiple Testing: Examples and Methods for *P*-Value Adjustmentxamples and Methods for *P*-Value Adjustment. John Wiley & Sons.
Westskog, H., Winther, T., Saele, H., 2015. The effects of in-home displays: Revisiting the contextevisiting the context. Sustainability 7, 5431–5451.
Zangheri, P., Serrenho, T., Bertoldi, P., 2019. Energy savings from feedback systems: A meta-studies' review. Energies 12 (19), 3788.