Discussion Paper Series – CRC TR 224

Discussion Paper No. 182
Project B 07

Disaggregate Consumption Feedback

Andreas Gerster[1]
Mark A. Andor[2]
Lorenz Goette[3]

May 2025
(First version: June 2020)

[1]University of Mainz, E-Mail: gerster@uni-mainz.de
[2]RWI – Leibniz Institute for Economic Research, Ruhr University Bochum
[3]National University of Singapore

Funding by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) through CRC TR 224 is gratefully acknowledged.

Collaborative Research Center Transregio 224 - www.crctr224.de
Rheinische Friedrich-Wilhelms-Universität Bonn - Universität Mannheim

Disaggregate Consumption Feedback

Andreas Gerster[*], Mark A. Andor[†], Lorenz Goette[‡]

May 12, 2025

Abstract

We investigate the impact of providing households with disaggregate consumption feedback and develop a framework to assess its welfare implications. In the context of smart metering, we find that the provision of appliance-level feedback causes an energy conservation effect of 5 percent relative to a group receiving standard (aggregate) feedback. Hence, a smart meter roll-out will be substantially more effective if appliance-level feedback is provided. We also show that the current regulatory approach to assess consumer surplus overestimates the gains from smart meter feedback.

JEL codes: D12, D83, Q41, Q58.

Keywords: Randomized controlled trial, disaggregation, consumption feedback, energy conservation.

Correspondence: Andreas Gerster, University of Mainz, Jakob-Welder-Weg 4, 55128 Mainz, Germany; Tel.: +49 6131 39 31551, E-Mail: gerster@uni-mainz.de.

Acknowledgements: We gratefully acknowledge financial support by the Swiss National Science Foundation (SNF) under grant 407140_153675 (NFP 71). Furthermore, this work has been partly supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) through CRC TR 224 (Project B07) and through CRC TRR 391 "Spatio-temporal Statistics for the Transition of Energy and Transport" (520388526). We are grateful for valuable comments and suggestions by Sarah Darby, Ken Gillingham, Timo Goeschl, Sébastien Houde, Katrina Jessoe, Andreas Lange, Mike Price, David Rapson, Ulrich J. Wagner as well as by participants of the 8th Atlantic Workshop, the 24th Annual Conference of the EAERE, the 1st RWI EEE Workshop, and seminars at RWTH Aachen, Technical University Berlin, UC Davis, University of Hamburg, University of Mannheim, Toulouse School of Economics, and Yale University. We thank Thomas Jaletzky, Gerhard Radtke, Lara-Franziska Rogowski, and Peter Schulten for excellent cooperation in conducting the field experiment as well as Kathrin Kaestner, Yero Ndiaye, Shawn Ho, Dorothy Ting, Chenrui Wang, and Yiwei Fan for excellent research assistance.

[*]University of Mainz and RWI – Leibniz Institute for Economic Research
[†]RWI – Leibniz Institute for Economic Research and Ruhr University Bochum
[‡]National University of Singapore

1 Introduction

Novel information technologies enable consumers to make better-informed decisions. For example, navigation systems assist car drivers in improving their travel planning (Chorus et al., 2006), text message reminders increase patients' adherence to medical treatments (Pop-Eleches et al., 2011), and fitness trackers help athletes to maintain higher levels of physical activity (Cadmus-Bertram et al., 2015). A key advantage of modern information technologies is their ability to provide disaggregate, i.e., behaviour-specific, feedback on the consequences of choice alternatives that are otherwise difficult to assess. For example, consumers have biased beliefs about the caloric content of food (Bollinger et al., 2011) and are unaware of the returns to education at different schools (Jensen, 2010), which compromises their ability to choose a healthy diet and a good school for their children. Providing disaggregate feedback solves these problems by informing consumers about the relative benefits and costs of their choice alternatives. Despite the rapid proliferation of information technologies, there is only limited evidence on the effectiveness of disaggregate feedback. In addition, a methodology for quantifying its benefits is unavailable so far.

In this paper, we explore the potential of disaggregate feedback in the context of smart metering and develop a framework to investigate the impact of feedback on consumer surplus. Governments throughout the world have implemented a massive deployment of advanced electricity metering infrastructure, which involves multi-billion investments into so-called smart electricity meters.[1] A core rationale for the deployment of smart meters is that these devices can provide households with consumption feedback. Feedback may foster awareness about the cost and environmental impact of electricity usage, and hence lead to energy conservation (EC, 2014b). It is well-documented that – in the absence of feedback – individuals tend to underestimate the energy use of energy-intensive appliances, and overestimate the energy use of low-intensity applications (Attari et al., 2010). Beyond exploring the impact of feedback on households' electricity-use behaviours, we also develop a method to assess its implications for consumer surplus. We derive sufficient statistics to quantify it, thus providing a novel tool for regulatory cost-benefit analysis.

[1] For example, 79 million households in the United States and 472 million households in China have been outfitted with smart meters by 2017 (IEA, 2017). In addition, the European Union has committed to installing 200 million smart meters at an estimated cost of 45 billion Euro (EC, 2014a).

We conduct a field experiment with 700 participants and provide electricity use feedback through a smartphone app. Participants in the Aggregate Feedback group obtain household-level feedback, as traditionally provided by smart metering interventions. Participants in our Disaggregate Feedback group additionally receive information on their appliance-level electricity use. Our study builds on a smart meter technology that leverages the potential of appliance-level feedback without installing additional costly infrastructure. The technology exploits that appliances leave distinct signatures in high-frequency aggregate electricity consumption data, which can be used to infer appliance-level uses. For our study, we collaborate with an electricity utility and use a product that has been validated for its accuracy in the process of its market introduction (for technical background information, see Gupta et al. 2010; Hart 1992; Gupta et al. 2017, and Appendix A3). For evaluating the impact of aggregate feedback, we construct a matched (non-experimental) control group of households of the same utility that have smart meters, but did not obtain any feedback. This allows us to establish a benchmark to a group without any feedback at all, which we exploit to identify the overall conservation effect.

We complement our field experiment with a theoretical model of household service demand, which serves two purposes. First, it enables us to derive sufficient statistics for evaluating the impact of feedback on consumer surplus. We also show how these statistics can be identified as simple functions of the aggregate and appliance-level treatment effects that we identify in our field experiment. Second, our model allows us to derive behavioural predictions regarding the effectiveness of aggregate and disaggregate feedback.

Our empirical results show that the provision of appliance-level feedback reduces electricity consumption strongly by around 5%, compared to our experimental condition in which individuals receive 'standard' aggregate smart metering feedback. Furthermore, we estimate that aggregate smart meter feedback reduces electricity consumption only by 1%.[2] Taken together, our estimates demonstrate that providing disaggregate information adds considerably to the effectiveness of feedback. Further evidence from secondary analyses suggests that households are poorly informed about the wattage of their appliances prior to our interven-

[2]This finding is in line with previous studies on aggregate feedback that have estimated conservation effects ranging from 2% to 3% (Carroll et al., 2014; Degen et al., 2013; Martin and Rivers, 2017), up to 5% (Schleich et al., 2017; Houde et al., 2013). Further studies have shown that feedback via more frequent billing can even lead to an increase in resource use (Wichman, 2017). With respect to consumption feedback, Gosnell et al. (2019) find that providing information on social comparisons and demand disaggregation via an advanced app is more effective than providing aggregate information via in-home displays and a basic app, but only for one of two smart meter installers. For reviews on feedback interventions, see e.g. Darby (2006) and Karlin et al. (2015).

tion, but improve the accuracy of their beliefs after receiving disaggregate feedback, in line with our theoretical framework.

We derive formulas to quantify consumer surplus that are simple variants of those currently used in regulatory cost benefit assessments. The changes in consumer surplus from disaggregate feedback can be calculated as the *weighted* sum over appliance-level cost savings. The weights are determined by consumers' relative bias, i.e., the perceived energy intensity divided by the actual intensity. In our study, we find that the disaggregate feedback increases consumer surplus by 5.4 EUR. We also derive bounds of the consumer surplus gain from aggregate feedback and quantify these bounds using our sufficient statistics. Our bounds imply that the welfare gains from aggregate feedback lie between 0.44 to 3.29 EUR per household and year, and are thus substantially smaller than those from disaggregate feedback.

We make two main contributions to the literature. First, we add to a literature on feedback by disentangling the effects of disaggregate appliance-level information from the effects of aggregate information. That households are only poorly informed about the cost of behaviours has been documented both in the context of energy use (Attari et al., 2010) and water use (Brent and Ward, 2019). Regarding the effectiveness of disaggregate feedback, the evidence is mixed. While Burkhardt et al. (2019) find no conservation effects of appliance-level feedback provided through an online portal, feedback on the energy use of showering, for example, has been shown to reduce water and electricity use by 10-20% (Asensio and Delmas, 2015; Bruelisauer et al., 2018; Tiefenbeck et al., 2018, 2019).

One difficulty in interpreting the evidence is that previous studies on appliance-level feedback have typically evaluated bundled interventions that also increase the salience of electricity use, reduce aggregate biases and improve energy-related knowledge, for example (Asensio and Delmas, 2015; Burkhardt et al., 2019; Tiefenbeck et al., 2018, 2019). Hence, it has proven difficult to assess the extent to which appliance-level feedback contributes to the overall effectiveness of an intervention. To isolate its contribution, we conduct a tailored randomized controlled trial that allows us to identify the additional savings induced by appliance-level information.

Our findings demonstrate that augmenting traditional smart meter feedback by appliance-level information could largely increase the effectiveness of a smart meter roll-out. So far, smart meters typically provide information about household-level electricity consumption. Previous evidence suggests that such feedback leads to modest conservation effects of 2% to maximally 5% (Degen et al., 2013; McKerracher and Torriti, 2013; Schleich et al., 2017; Houde et al., 2013). Against this background, our finding of an additional conservation effect of 5% from appliance-level feedback suggests that such feedback is crucial for effective smart meter interventions.

Second, our study relates to the literature on welfare analysis under optimization errors by consumers (see Farhi and Gabaix 2020 for an overview). One strand of this literature has primarily focused on the implications of tax misperceptions (e.g., Chetty et al. 2009; Rees-Jones and Taubinsky 2019). Another strand has derived optimal corrective taxes and subsidies for behaviourally biased consumers (e.g., Allcott and Taubinsky 2015; Gerster and Kramm 2019; O'Donoghue and Rabin 2006). By contrast, we analyse informational instruments when consumers misperceive product attributes and the cost of household services. We derive formulas for evaluating consumer surplus that can be used for policy analysis of informational interventions to overcome such misperceptions.

Our findings demonstrate that the current regulatory approach to assess consumer surplus gains from smart meter feedback is fundamentally flawed. Cost-benefit analyses in the U.S. and the EU, for instance, approximate changes in consumer surplus by the realized energy cost savings (Faruqui et al., 2011; Giordano et al., 2012). In our study, disaggregate feedback reduces expenditures by 48.3 EUR on average, while consumer surplus increases only by 5.4 EUR. Hence, consumer surplus gains are overestimated by a factor of about ten. This deviation arises because the weights used to calculate consumer surplus are substantially smaller than one (in absolute terms). Intuitively, a reduction of electricity consumption in response to feedback involves not only a financial gain from lower expenditures, but also a utility loss from consuming less of a household service, which is neglected in current cost-benefit analyses.

The remainder of the paper is structured as follows. Section 2 introduces our model. Section 3 presents the experimental design and the data. In Section 4, we estimate the impact of aggregate and appliance-level feedback on energy consumption and present results from secondary analyses. Section 5 quantifies the impact of appliance-level feedback on consumer surplus. Section 6 concludes.

2 Conceptual Model

We start by investigating the effects of providing appliance-level feedback to consumers based on the Becker (1965) household services model. Let consumers have the following quasi-linear utility function:

$$U(\mathbf{x}, z) = u(\mathbf{x}) + z,$$

where $u(\mathbf{x})$ is quasi-concave and denotes the utility from consuming J household services denoted by the vector $\mathbf{x} = (x_1, \ldots, x_J)'$ and z represents the numeraire good, whose price is normalized to 1. The consumption of household service j requires inputs of $y_j = x_j e_j$, where e_j denotes the input intensity of service x_j, which by definition is non-negative. In our application, households consume energy services by using a particular appliance, such as a dish-washer or dryer, and the input intensity refers to the amount of electricity that is needed to operate an appliance. Consumers maximize their utility subject to the budget constraint $w = z + \sum_j y_j p$, where w denotes their exogenous income and p denotes the price of the input, in our case electricity.

In line with the literature (e.g., Attari et al. 2010), let consumers have biased perceptions of input intensities $\tilde{e}_j = e_j + b_j$, where b_j denotes a bias term. We decompose the bias further as $b_j = b^a + b^s_j$, where b^a denotes an aggregate bias term and b^s_j denotes a service-specific bias that affects only those beliefs regarding appliance j.

2.1 Behavioural Predictions

We now derive predictions of providing aggregate and disaggregate, i.e., service-specific, feedback on total input use. In line with the current literature, we assume that aggregate and disaggregate feedback entirely remove aggregate and service-specific bias, respectively (Chetty et al., 2009; Bernheim and Taubinsky, 2018).

Under a general utility function $u(\mathbf{x})$, we find correcting a bias of underestimating aggregate service intensities reduces the total amount of input use for service consumption if all services are a normal good (see proposition 1 in Appendix A1.1 for a derivation). Intuitively, a reduction in total input use arises because correcting an underestimation increases the perceived cost of using any appliance, which reduces the input demand for any appliance and, hence, total input demand. For the case of household electricity consumption, studies have consistently found small but negative demand elasticities (see, e.g., Frondel et al., 2019). Thus, we expect aggregate consumption to fall with aggregate feedback.

In a next step, we explore the impact of providing disaggregate feedback in addition to aggregate feedback in order to remove the service-specific biases. To derive a prediction, we need to impose some structure on the shape of service-specific biases. Empirical studies from the domain of energy consumption show that individuals tend to overestimate the energy intensity of low-intensity appliances, while underestimating the energy intensity of high-intensity appliances (Attari et al., 2010; Fang et al., forthcoming). Hence, we model perceived energy intensity towards the mean: $\tilde{e}_j = \alpha e_j + (1-\alpha)e$, where $\alpha \in [0,1)$ is the weight attached to the service intensity of service j, and $e = \sum_j g_j e_j$ is the usage weighted average service intensity.[3] In addition to empirical realism, this formulation has a straightforward psychological interpretation: the weight α can be interpreted as the outcome of a cognitively uncertain updating from signals about each services' intensity (Enke and Graeber, 2023; Gabaix, 2017).

This specification implies that eliminating disaggregate biases may increase input intensity beliefs for some appliances and reduce them for others. It is therefore not obvious under what conditions removing the service-specific bias will lead to a decrease in overall input use. In proposition 2 in Appendix A1.2, we show that overall input use decreases if the price elasticity of the more energy intensive service is at least as high as the price elasticity of the less energy intensive service. The rationale is as follows: even though disaggregate feedback may induce participants to increase the energy use of the less energy intensive service, this increase is overcompensated by the energy savings from learning about the high energy intensity of the other service. Whether such a decrease in total input use materializes in practice is ultimately an empirical question, which we explore in the empirical part of this paper.

Our discussion also clarifies that disaggregate feedback need not cover all input uses to be effective. To fix ideas, let a customer not only receive aggregate feedback, but also disaggregate feedback for one of two household appliances. In that case, feedback conveys all information about the other appliance, which can be derived from the difference between the aggregate feedback and the disaggregate feedback on the one appliance covered. For cases with more than two appliances, consumers also learn about the service without disaggregate feedback, but not completely. In particular, they can learn about the sum of the appliances without

[3]Notice that this formulation satisfies our mean-zero condition for service specific biases, as $\sum g_j \tilde{e}_j = \alpha \underbrace{\sum g_j e_j}_{\equiv e} + (1-\alpha)e = e$

disaggregate feedback by assessing by how much the aggregated feedback exceeds the sum of all appliances, for which disaggregate feedback is provided. Thus, even though disaggregate feedback may be feasible only for large appliances, it allows consumers to learn about smaller appliances as well.

2.2 Consumer Surplus

For our welfare analyses, we assume that utility is additively separable, but instead do not impose any structure on the magnitude of aggregate and disaggregate biases.[4] Furthermore, in line with the literature, we assume that aggregate and service-specific feedback fully eliminates the respective bias (see, e.g., Allcott and Taubinsky 2015).[5] The focus of our approach is to identify the welfare gain of service-specific feedback over and above what aggregate feedback can achieve. In Appendix A2.1, we derive that a second-order approximation of the change in consumer surplus in response to an intervention that removes aggregate and disaggregate biases is given by:

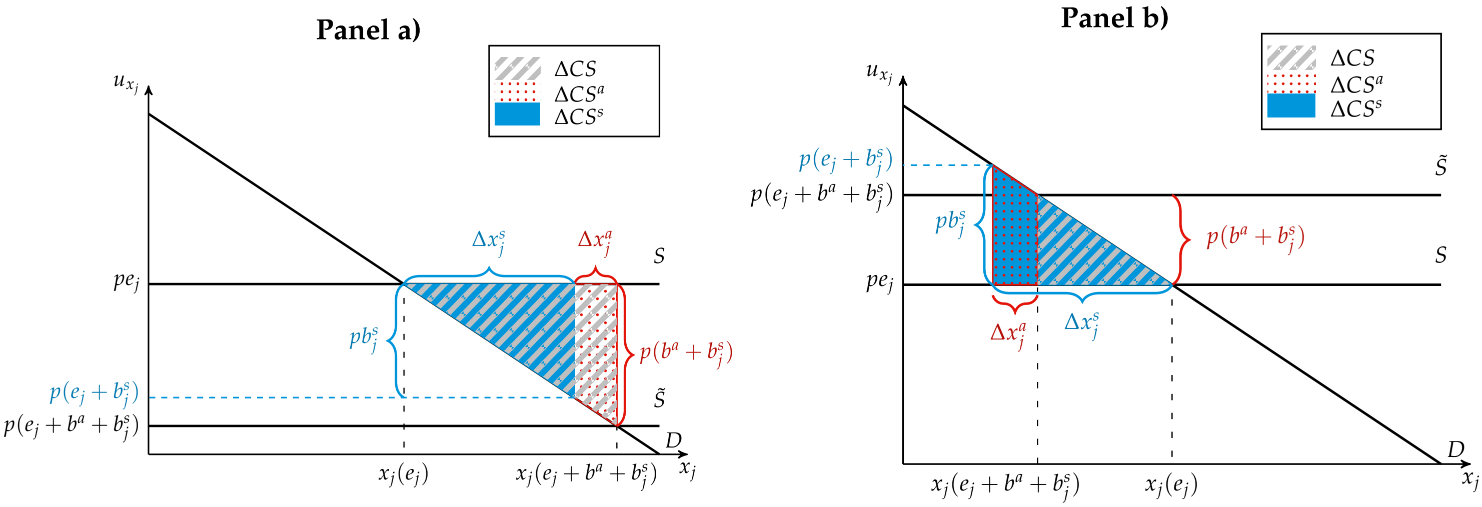
$$\begin{aligned} \Delta CS &= \Delta CS^a(b^a = 0, \mathbf{b^s}) + \Delta CS^s(b^a = 0, \mathbf{b^s} = \mathbf{0}) \\ &= \sum_j \left(\frac{b_j^s}{e_j} + \frac{b^a}{2e_j} \right) \Delta E_j^a + \sum_j \frac{b_j^s}{2e_j} \Delta E_j^s, \end{aligned} \tag{1}$$

where ΔCS denotes the change in consumer surplus, and ΔE_j^a and ΔE_j^s are the average treatment effects on the input expenditures for household service j in response to an elimination of aggregate and service-specific biases, respectively. We partition the calculation of the overall gain in consumer surplus into two steps. First, we determine the consumer surplus change from removing the aggregate bias b^a, $\Delta CS^a(b^a = 0, \mathbf{b^s})$, while leaving the service-specific biases $\mathbf{b^s} = (b_1^s, b_2^s, ..., b_n^s)$ in place. In the context of electricity consumption, this term refers to the gain in consumer surplus of a conventional smart metering intervention that does not provide appliance-specific feedback. In a second conceptual step, we then additionally remove the service-specific biases $\mathbf{b^s}$, as reflected in the term $\Delta CS^s(b^a = 0, \mathbf{b^s} = \mathbf{0})$.

[4] As we show in Appendix A1.2, one could alternatively maintain the general formulation of $u(\mathbf{x})$ and the specific form of bias we assume above. The appendix derives the sufficient statistics for welfare analysis in this case, which may be more appropriate for other contexts.

[5] If the elimination of biases is only partial, our calculations will underestimate the welfare effects. Our approach assumes that, for a given behavioural response that enters the calculations, bias is entirely eliminated. Thus, the last unit of change in consumption will increase consumer surplus by approximately zero. However, if the bias is not entirely removed, that last unit still produces first-order gains in consumer surplus.

Figure 1: Changes in Consumer Surplus in Response to Aggregate and Disaggregate Feedback



Notes: Panel a) depicts a situation where the aggregate bias, b_a, and service-specific bias, b^s_j, are both negative. Panel b) depicts a situation with the same aggregate bias, but a positive service-specific bias. ΔCS refers to the overall consumer surplus change, ΔCS^a refers to the change from removing an aggregate bias, and ΔCS^s refers to the change removing the service-specific bias (given that aggregate biases have been removed already).

Note first that each summand of the term $\Delta CS^s(b^a = 0, \mathbf{b^s} = \mathbf{0})$ is positive. If $b^s_j < 0$, i.e., if the individual underestimates the service intensity e_j, then service-specific feedback will reduce the consumption of and the expenditures on x_j. Hence, $b^s_j \Delta E^s_j < 0$ is positive for every service j. Conversely, if $b^s_j > 0$, service-specific feedback will lead the individual to realize that e_j is lower than she thought, which implies that $\Delta E^s_j > 0$. Both cases are illustrated in Figure 1 Panel a) and b), respectively, where the blue-shaded areas correspond to $\Delta CS^s(b^a = 0, \mathbf{b^s} = 0)$. The key finding is that providing service-specific feedback on top of aggregate feedback always increases consumer surplus, as intuition would suggest. As we show in Appendix A2.3, this finding holds under very general assumptions, e.g. also in the case of non-separable utility.

Consider now the term $\Delta CS^a(b^a = 0, \mathbf{b^s})$, as depicted in Panel a) of Figure 1. If both b^a and b^s_j go in the same direction, then removing the aggregate bias b^a increases consumer surplus by the area shaded in red. Panel b) in Figure 1 displays the case where the aggregate and service-specific bias go in opposite directions. In this case, removing aggregate bias alone can decrease consumer surplus, as the example shows. The removal of b^a leads to a larger discrepancy between perceived and actual service intensity, and thus distorts choices even more. Thus, while removing all biases always unambiguously increases consumer surplus, removing only one component of a bias may well harm consumers.[6]

[6]The result that opposing biases mitigate losses in consumer surplus is well-known (see, e.g., Benabou and Tirole (2002)).

Equation (1) highlights three major flaws of current cost-benefit analyses that approximate changes in consumer surplus by the sum of expenditure savings (e.g., Faruqui et al. 2011; Giordano et al. 2012). First, the change in consumer surplus equals the *weighted* sum of changes in expenditures, where the weights are given by the relative aggregate and disaggregate biases. These weights are typically less than one in absolute value, and thus reduce consumer surplus estimates below the levels used in current cost-benefit analyses. Second, expenditure savings are not a necessary condition for increases in consumer surplus, as implicitly assumed in those analyses. If consumers overestimate the cost of using an energy service j, their welfare increases if they use that service more, thereby increasing expenditures. Third, the relationship between consumer surplus and expenditures may break down entirely when only aggregate feedback is provided. To see that, assume that Panel a) and Panel b) refer to two distinct appliances used by a consumer. Since the consumer will use both appliances less in response to aggregate feedback, the effect on expenditures is unambiguously negative. However, the consumer will incur welfare losses from the use of the appliance depicted in Panel b), for which she overestimates the energy intensity. Overall, it may be even possible that these welfare losses dominate the welfare gains from the reduction of the use of the other appliance, although electricity consumption and expenditure is reduced. Hence, inferring even directional changes in consumer surplus from changes in expenditures is infeasible in general.

2.3 Sufficient Statistics

Equation (1) provides us with the structure to estimate the impact of feedback on consumer surplus based on few statistics. In addition to the readily observable appliance-specific treatment effects of aggregate feedback on expenditures (ΔE_j^s and ΔE_j^a), we need to quantify the relative bias terms (b_j^s/e_j and b^a/e_j).

The following intuition guides our identification (see also Chetty, 2009): appliance-specific feedback changes the perceived prices of service j by $-b_j^s/e_j$ percent. The treatment effect of appliance-specific feedback $\Delta y_j^s/y_j$ thus measures the percent response to this perceived price change. The demand elasticity η_j of service j indicates the percent change of y_j in response to a one-percent increase in the price. Thus, we can infer the relative service-specific bias from the relative treatment effect of appliance-specific feedback, normalized by the demand elasticity, $b_j^s/e_j = -(\Delta y_j^s/y_j)/\eta_j$. As we show in Appendix A2.1, this yields

$$\Delta CS^s(b^a = 0, \mathbf{b^s}) = -\sum_j \frac{\Delta y_j^s/y_j}{2\eta_j} \Delta E_j^s. \tag{2}$$

Our empirical setup does not allow us to point-identify $\Delta CS^a(b^a = 0, \mathbf{b^s} = \mathbf{0})$. The reason is that we would also need to observe appliance-specific consumption in the matched control group. However, we only observe aggregate consumption in that group. Yet, we are able to bound the change in consumer surplus from aggregate feedback by:

$$-\left(\frac{\Delta y_n^s/y_n}{\eta_n} + \frac{\Delta y^a/y_n}{2\eta_n}\right)\Delta E^a \leq \Delta CS^a(b^a = 0, \mathbf{b^s}) \leq -\left(\frac{\Delta y_m^s/y_m}{\eta_m} + \frac{\Delta y^a/y_m}{2\eta_m}\right)\Delta E^a, \tag{3}$$

where Δy^a and ΔE^a denotes the ATE of an intervention that removes aggregate biases on input expenditures and use, respectively. Furthermore, m and n denote the service j with the maximum and minimum relative bias under full attribution of the aggregate effect size Δy^a to that service, respectively.[7]

All sufficient statistics to quantify Equations (2) and (3) are identified by our study design. To obtain estimates for price elasticities, we exploit cross-sectional price variation across German regions to estimate them (details are provided in Section 5). This variation partly stems from differences in grid surcharges, which are higher in regions with substantial electricity generation from renewable energy sources. For sensitivity analyses on the role of our elasticity estimates, we also use aggregate estimates from the literature.

3 Experimental Design and Data

3.1 Study Groups

We draw on data from two populations: an experimental sample that was recruited for the study and randomized into conditions receiving aggregate or disaggregate feedback, and a population of households with pre-installed smart meters who receive no feedback at all. The information treatments of the experimental sample are provided via an app that study participants install on their smartphones and tablet PCs. After an initial login, participants can

[7]In mathematical terms: $m = max_j \tilde{CS}_j^a = -\left(\frac{\Delta y_j^s/y_j}{\eta_j} + \frac{\Delta y^a/y_j}{2\eta_j}\right)\Delta E^a$, and, equivalently $n = min_j \tilde{CS}_j^a$.

access the app without entering their credentials, except if they deliberately log out. Feedback is provided through the app and its specific functionalities were determined as part of the experiment.

We randomly assigned participants into one of two experimental treatment conditions. Participants in our Aggregate Feedback (A) group get access to an app that provides information about their household-level electricity use. On the start screen, participants can observe a real-time power meter that visualizes their current wattage. They also see the cost of their current monthly electricity consumption relative to their monthly advance payment.[8] On an additional screen, participants can compare their electricity consumption with their own history, as well as with other study participants, at monthly, weekly, daily, and hourly frequencies. They may also earn electronic 'badges' for completing their personal profile on the app and take part in an energy-related quiz. Participants in our Disaggregate Feedback group (D) have access to the same app, but can use an additional functionality that provides feedback on appliance-level usages (for screenshots of the smartphone app in each experimental condition, see Figures A2, A3, and A4 in the Appendix). All other functionalities and defaults are identical for both groups.[9]

In addition to the experimental sample, we obtain data from a non-experimental sample of households with smart meters that are served by the same utility. These households have agreed to report their electricity consumption to the grid operator who uses the data to forecast load profiles, but receive no feedback on their electricity use. We obtain smart meter data in 15 minute intervals for 577 households, starting from November 1, 2016, which is when our field test started. We also obtain data on the last annual bill prior to that date, as for our experimental sample. Using a 1-to-1 propensity score matching procedure, we construct a Matched Control (MC) group for the aggregate feedback group. We use propensity score matching on baseline consumption and the billing cycle dates to minimize differences between the two groups (for details, see Appendix A4). The MC group serves as a benchmark to identify the conservation effects of aggregate feedback, as discussed in detail in Section 3.4.

[8] In Germany, typically, billing occurs annually and monthly advance payments are intended to smooth electricity costs over the year. Exceeding the monthly advance payment has no financial consequence, but indicates that households may face additional payments when the next yearly billing occurs.

[9] We further subdivided the disaggregation group into four treatments, with the aim of strengthening the appliance-level feedback, by introducing social comparison and additional financial incentives. These treatments are secondary to the goal of this paper and are discussed in Appendix A5.5.

3.2 Study Implementation

We conducted the randomized controlled trial with customers of a large German utility. Customers were invited to take part in a smart meter study via an email that did not mention appliance-level feedback as the purpose of the study. To be eligible for participation in the experiment, customers had to have a smartphone and wireless internet access. Furthermore, participants who own solar photovoltaic panels were excluded. Out of around 50,000 customers we invited, 800 participants agreed to take part in the study and met our eligibility criteria. All participants have a two-part electricity tariff, which is common in Germany. They pay a flat rate for every kilowatt hour of electricity consumed and a fixed annual base price.

All participating households received a high-resolution smart meter, an internet gateway that connected the smart meter with the internet, and access to a smartphone app. After the smart meter had been installed by professionals, the utility sent participants the internet gateway along with instructions how to install the app. As soon as participants had activated the gateway and installed the app, they shared their smart meter data and our study started.[10]

More than 90% of our study participants entered the field test between November, 2016, and January, 2017, and the remaining participants joined afterwards. The core study period extended for 6 months, when consumers had access to the full functionality of the app in their respective treatment group. From month 7 onwards, households were free to continue to use the app for another three months. As the number of participants declined considerably during that period, our analysis focuses on the core study period.

The smart meters measure electricity consumption at a high frequency, typically every second. This results in a rich dataset of several billion observations over the entire study period. The high granularity of our data allows us to use commercial load disaggregation techniques that disentangle the total electricity use into appliance-level uses.[11] The smart meter data is saved online in real-time and processed daily to detect appliance usages based on a so-called nonintrusive appliance load monitoring (NALM) algorithm, which employs machine learning techniques for load disaggregation. The algorithm exploits the fact that appliances have characteristic electricity use signatures. These signatures can be used to disaggregate

[10]In this setting, the utility deemed a pure experimental control group as infeasible and feared the confusion of participants that participated in the trial, got a smart meter installed, but no visualization of smart meter data.

[11]This feature sets our study apart from earlier attempts to analyse and disaggregate smart meter data, such as Google PowerMeter or Microsoft Hohm, which were discontinued in 2011 and 2012, respectively. A main reason for the failure of these services was insufficient access high-granular smart meter data (Donnal and Leeb, 2015). The high accuracy of the algorithm we use in this study has been documented in verification studies, which have shown that it detects 94% of all appliance-level uses (Gupta et al., 2010; Carrie Armel et al., 2013).

Table 1: Descriptives

	Matched Control (MC)	Agg. Feedback (A)	Disagg. Feedback (D)	P-value
Baseline consumption, in kWh/day	10.0	10.2	10.4	0.65
No. of occupants	–	2.5	2.6	0.32
Monthly net income, in EUR	–	3,004	3,127	0.32
Own property, in %	–	74.3	75.5	0.76
Employed, in %	–	50.2	51.4	0.73
Share of females, in %	–	44.8	48.0	0.09
Age, in years	–	47.6	45.4	0.25
Number of households	140	140	560	$\sum$=840

Notes: P-values are from F-tests of mean equality between households in the Aggregate Feedback (A) and Disaggregate Feedback (D) groups, clustered at the household level. Socio-demographics are not available for participants in the Matched Control group (MC). For groups A and D, they are measured at the household level, except for employed, share of females, and age, which we measure at the household member level.

high-resolution smart meter data into appliance-specific electricity uses (see Appendix A3 for details). Detection of appliance-level uses is possible for the major appliances of a typical household, including the categories *Dishwasher*, *Washing Machine*, *Dryer*, and *Oven*, as well as a *Refrigeration* category that captures refrigerators and freezers. The algorithm also identifies an *Always-on* category as the typical consumption at 3 a.m. In addition to the appliance categories that we can directly measure, we construct a residual category *Other Appliances*, which captures the electricity consumptions of all other appliances.

Table A6 in the Appendix gives descriptive statistics on the more than 300,000 appliance events that we observe during the core study period. To test the plausibility of the appliance-level measurements, we compare them to typical appliance-level uses in Germany, which are available for 2006 and 2011. We find only small differences for refrigeration and residual use, which likely reflect improvements in energy efficiency in refrigeration and a general trend that households use more electric devices (see Appendix A3.2 for details).

3.3 Data

Of the 800 participants in our experimental sample, information on the electricity use in the annual billing period prior to the field test is missing for 27 participants.[12] Furthermore, 73 households experienced technical difficulties that prevented them from connecting their smart

[12] For 18 participants, baseline electricity use is missing. In addition, we set electricity baseline to missing when the difference between the baseline and the experimental period is in the top or bottom percentile of its distribution within each treatment group (e.g. above +126% and -64% for group A) or larger than 25 kWh per day in absolute terms, which concerns 9 participants.

meter to the internet. As a result, the final experimental sample used for our analyses consists of 700 participants. The number of participants in the Disaggregate Feedback group is higher than the number in the Aggregate Feedback group (560 vs. 140), because we implement three additional sub-treatment arms that test whether financial incentives or social information can increase the engagement with disaggregate information. Because we find that this is not the case, we relegate the discussion of these sub-treatments to Section 4.3. As a consequence of 1:1 matching, our matched control group consists of 140 participants, as the Aggregate Feedback group. For all households, we observe smart meter data during the intervention period, as well as their most recent annual electricity use before the field test started, which serves as our baseline. For households in our experimental sample, we additionally conducted surveys to elicit participants' socio-demographic characteristics, attitudes, and beliefs.

In Table 1, we show that socio-demographic variables and the electricity use during the baseline period are balanced across our experimental groups, as expected from randomization. When we test for mean equality across our experimental conditions, we cannot reject the null hypothesis that the variable means are equal at the 5% level for any covariate, as shown in the last column of Table 1. The same holds true for various dwelling characteristics and the possession of household appliances such as cooling appliances, washing machines or tumble dryers (for details, see Appendix Table A3). Hence, we can rule out the possibility that participants in our experimental groups differ systematically in their household equipment, which might otherwise confound our treatment effect estimates.

As shown by Table A2 in the Appendix, socio-demographic characteristics of study participants are comparable to German averages. In terms of age, employment status, sex, and household net income, our experimental sample is similar to the German population. Participating households consist of slightly more occupants (2.5 vs. 2.0 in Germany), which is mostly driven by a smaller percentage of single-person households (12% vs. 42% in Germany). For households of a given size, electricity consumption levels are similar in our study and the German population. The larger average number of occupants per household in our sample translates into larger average electricity consumption levels compared to the German average (10.4 vs. 8.6 kWh per day). Furthermore, households in our sample live more often in their own property than the average German household (76% vs. 44% in Germany) and went to school for slightly longer (11 vs. 10.5 years).

3.4 Empirical Strategy

The main goal of our randomized controlled trial is to identify the average treatment effect (ATE) of providing disaggregate feedback in addition to aggregate feedback. Furthermore, we exploit the group MC to estimate the ATE of aggregate feedback on electricity consumption, and to obtain an estimate of the total electricity savings from disaggregate feedback relative to a group without any feedback.

Our design allows us to experimentally identify the impact of disaggregate feedback. Randomization ensures that participants in group A and D are identical in terms of both observable and unobservable characteristics (see Table 1 and Appendix Table A3 for the absence of statistically significant pre-treatment differences). We use participants in the group A and D to estimate the equation:

$$Y_{it}^{norm} = \alpha Y_i^b + \beta D_i + \nu_t + \mu_m^b + \epsilon_{it}, \tag{4}$$

where D_i is a dummy variable that equals one if a household received disaggregate feedback and zero otherwise. The variable Y_{it}^{norm} denotes electricity use of household i at day t, divided by the average daily electricity use in the aggregate feedback group during the core study period.[13] The variable Y_i^b denotes the average daily consumption during the baseline period, normalized the same way. Including it in our equation allows us to control for permanent between-household differences in electricity consumption. This mimics a difference-in-difference design and increases the efficiency of our estimates. We also control for billing cycles by including a set of month-of-baseline fixed effects (μ_m^b) that equal one if the baseline metering period ended in month m, and zero otherwise. Our model includes day fixed effects (ν_t), which absorb variation from seasonality, and an error term ϵ_{it}. We cluster standard errors at the household level, thereby accounting for serial correlation in the error terms.

To identify the effect of providing aggregate feedback, and the effect of disaggregate feedback relative no feedback at all, we include participants from the group MC and estimate the following equation:

$$Y_{it}^{norm} = \alpha Y_i^b + \gamma A_i + \delta D_i + \nu_t + \mu_m^b + \epsilon_{it}, \tag{5}$$

[13] This normalization expresses treatment effects as a percentage of the average consumption level in the absence of treatment and is common in the literature (e.g., Allcott 2015). It also provides us with a direct link to the sufficient statistics in the welfare analysis. The treatment effects in kWh can be obtained by simply multiplying our estimates with the Aggregated Feedback (A) group mean (10.4 kWh).

where all the variable definitions are as in Equation (4), and A_i equals one if household i is in group A, and zero otherwise. Notice that the reference group in Equation (5) is the group MC that does not receive any feedback. Thus, γ measures the effect of receiving aggregate feedback alone, and δ measures the overall effect of disaggregate feedback, i.e. $\delta = \beta + \gamma$. We correct the standard errors for clustering at the household-match level, i.e., we assign the same clustering unit to matched households from the groups A and MC (Abadie and Spiess, 2021).

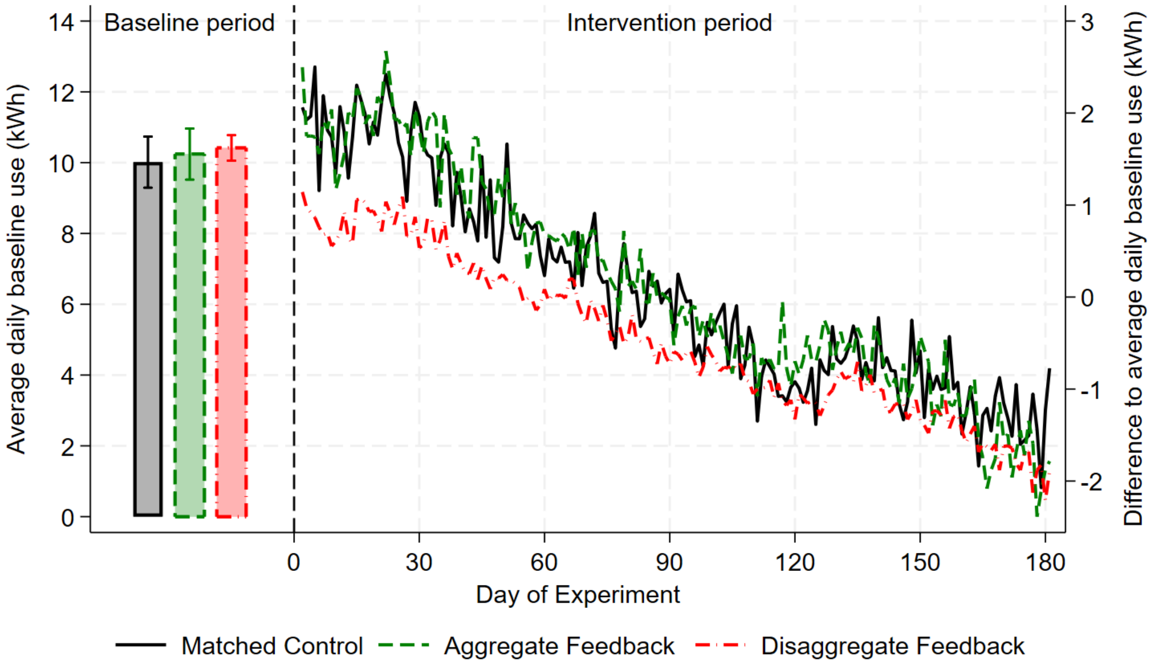
Consistent estimation of γ and δ requires a Conditional Independence Assumption (Imbens and Wooldridge, 2009): conditional on covariates, the treatment group indicators D and A need to be independent of the error term ϵ. Controlling for baseline use eliminates any bias from differences in levels. However, the Conditional Independence Assumption also requires trends in each group to be parallel in the absence of an intervention. While randomization ensures that this is the case for the identification of β, it could in principle be violated for γ. For example, a violation could arise if baseline electricity consumption levels were unbalanced across study group and if trends differed depending on these levels. In our setting, matching ensures that baseline consumption levels are balanced across study group, which mitigates such concerns. Furthermore, households could face different weather shocks during the outcome period, which could confound our estimates. In Appendix Figure A8, we plot average sunshine, precipitation and temperature in our study groups and find no empirical support for such concerns in the context of our study in Germany.

4 Results

4.1 Effect of Feedback on Total Electricity Use

We start by descriptively investigating the impact of aggregate and disaggregate feedback on electricity consumption. The right panel of Figure 2 shows the difference between the daily electricity use during the intervention period and the average daily use in the baseline period. The electricity use of households in group D have considerably lower average electricity use levels than households in the other study groups after our intervention begins, which is first evidence that disaggregate feedback reduces electricity consumption beyond aggregate feedback. By contrast, the average electricity use levels are very similar for the groups A and MC, which speaks against pronounced electricity conservation effects in response to receiving aggregate feedback. For all study groups, the average daily electricity use declines over time,

Figure 2: Average Daily Consumption by Study Group



Notes: The left part of the figure shows the average daily electricity use during the last annual billing period prior to the intervention for our three study groups. The right part plots the average daily consumption of electricity during the intervention period, demeaned by average billing baseline consumption.

which reflects that our study starts in winter, when electricity use tends to be the higher than in other seasons.

We continue by econometrically estimating the ATE of disaggregate feedback, relative to obtaining aggregate feedback. Exploiting experimental variation only, we estimate Equation (4), finding that the ATE amounts to -4.8% and is statistically significant at all conventional levels (Column 1 of Table 2). Hence, providing disaggregate feedback in addition to aggregate feedback yields large additional reductions in total electricity input. Our estimate remains virtually unchanged when we additionally control for weather controls such as sunshine, precipitation, and temperature, which provides further evidence that weather shocks do not confound our estimates.

In Column (4), we present the estimates of both aggregate and disaggregate feedback, relative to not obtaining any feedback, from estimating Equation (5). We find that the ATE for households who obtain aggregate feedback amounts to only -0.9%. One explanation is aggregate feedback may not be sufficient to overcome the search frictions that prevent consumers from acquiring information about which appliances might be responsible for high aggregate electricity consumption levels. Our estimate is consistent with previous studies that have found only small savings from aggregate feedback in European countries (e.g., Degen et al. 2013). We again find no evidence that weather shocks confound our estimates. While restricting the analysis to our subsample that can be georeferenced and linked to weather data reduces our sample size and slightly increases our point estimates (Column 5), additionally including weather controls leaves our estimates virtually unaffected (Column 6).

4.2 Effect of Disaggregate Feedback on Appliance-Level Electricity Use

We proceed by exploring the appliance-specific ATEs of disaggregate feedback, which we have identified as sufficient statistics for evaluating welfare effects. For that purpose, we estimate Equation (4) separately for every appliance category, substituting the outcome variable by the average daily consumption of each appliance category, normalized by the respective average in group A.

As shown in Panel b) of Table 2, we find that the conservation effects from appliance-level feedback are close to zero for appliance categories which are typically used throughout the day, such as *Refrigeration* and *Always-On*. The low response may partly reflect that consumers' demand for refrigeration is largely constant, irrespective of (perceived) cost. By contrast, we find that appliance-level feedback triggers a substantial reduction in the electricity consumption of dryers (Column 5 of Table 2b). As dryers are an electricity intensive appliance, it is plausible that consumers underestimate it (Attari et al., 2010) and hence reduce their energy consumption after receiving appliance-level feedback. In addition, substitutes for using the dryer are often available as dry-hanging clothes is common for German households. We also find some evidence that participants have reduced their use of the dish-washer, yet this effect is not statistically significant at any conventional level.

In addition, we find that households reduce consumption for the category *Other Appliances*, an effect that is statistically significant at the 1% level. This category encompasses a variety of electric appliances, such as televisions, hi-fi systems, vacuum cleaners, computers, as well as lighting. While no direct feedback is given for these appliances, our evidence suggests that participants nonetheless may have updated their respective beliefs. This is possible because more accurate feedback about some of the appliances helps them attribute the residual electricity use to the remaining devices. Our evidence suggests that they learned that these other

Table 2: ATEs on Daily Electricity Consumption

(a) ATE of Aggregate and Disaggregate Feedback (at the Household-Level)

	(1)	(2)	(3)	(4)	(5)	(6)
A: Aggregate Feedback	-	-	-	-0.009	-0.027	-0.029
				(0.022)	(0.029)	(0.029)
D: Disaggregate Feedback	-0.048***	-0.046***	-0.047***	-0.055***	-0.072***	-0.074***
	(0.016)	(0.018)	(0.018)	(0.016)	(0.023)	(0.023)
Sunshine (in min per hour)	-	-	-0.001**	-	-	-0.001**
			(0.000)			(0.000)
Precipitation (in liters per hour)	-	-	0.032**	-	-	0.029**
			(0.015)			(0.014)
Temperature (in °C)	-	-	-0.007**	-	-	-0.006**
			(0.003)			(0.003)
Day fixed effects (FE)	✓	✓	✓	✓	✓	✓
Month-of-baseline FE	✓	✓	✓	✓	✓	✓
Subsample with weather data		✓	✓		✓	✓
Y^b: Baseline elec. use	0.895***	0.904***	0.906***	0.910***	0.907***	0.908***
	(0.022)	(0.023)	(0.023)	(0.020)	(0.022)	(0.022)
R^2	0.5586	0.5684	0.5689	0.5687	0.5631	0.5636
Number of obs.	106,283	93,350	93,350	127,790	104,401	104,401
Number of participants	700	613	613	840	684	684

(b) ATE of Disaggregate Feedback (at the Appliance-Level)

	(1) Always-On	(2) Refrigeration	(3) Dish−Washer	(4) Washing	(5) Dryer	(6) Oven	(7) Other appl.
D: Disagg. Feedback	-0.002	-0.007	-0.091	-0.028	-0.439***	0.024	-0.072***
	(0.046)	(0.041)	(0.085)	(0.064)	(0.159)	(0.151)	(0.026)
Y^b: Baseline elec. use	1.156***	0.461***	0.747***	0.637***	1.064***	1.354***	0.851***
	(0.065)	(0.086)	(0.103)	(0.072)	(0.166)	(0.197)	(0.033)
Day fixed effects (FE)	✓	✓	✓	✓	✓	✓	✓
Month-of-baseline FE	✓	✓	✓	✓	✓	✓	✓
R^2	0.367	0.152	0.046	0.028	0.035	0.040	0.356
Number of obs.	93,187	93,185	84,511	91,473	65,852	93,187	93,187
Number of households	700	700	635	686	499	700	700

Notes for Panel a): ***, **,* denote statistical significance at the 1%, 5%, 10% level, respectively. Standard errors are in parentheses and clustered at the household (Columns 1-3) and household-match level (Columns 4-6), respectively. The outcome variable is daily electricity consumption, divided by the mean in the A group (10.4 kWh). A and D equal one for the households that obtain aggregate and disaggregate feedback, respectively.
Notes for Panel b): The outcome variable is daily electricity consumption of an appliance, divided by the mean for the same appliance in the A group (2.30, 0.97, 0.29, 0.47, 0.16, 0.20, and 5.6 kWh for Columns 1 to 7, respectively). The number of observations varies across columns as not all households possess all appliances. Standard errors are in parentheses and clustered at the household level.

devices were more energy-intensive than they originally thought, and thus cut back on usage. In Appendix A1.3, we show that this is indeed consistent with Bayesian updating by rational individuals.

4.3 Secondary Analyses and Robustness Checks

In the following, we briefly discuss the findings from a series of secondary analyses and robustness checks (for details see Appendix Section A5).

Treatment Effects by Baseline Use. We find that the treatment effects from disaggregate feedback are particularly large for households with high levels of baseline electricity consumption (see Subsection A5.1 and Table A9 in the Appendix).

Treatment Effects by Hour-of-the-Day. We show that treatments effects occur predominantly during daytime, which is when households typically use appliances (see Subsection A5.2 for details). The treatment effects of disaggregate feedback are particularly large during late morning hours and late evening hours, while we estimate that the savings from aggregate feedback occur during early evening hours.

Role of Beliefs. We conducted three surveys during the course of the study to analyze, among others, whether there are biased beliefs regarding aggregate electricity consumption and appliance-specific electricity consumption. In particular, we carried out the belief elicitations before and after the start of the interventions in order to estimate the effects of the interventions on the beliefs (see Subsection A5.4 for details).

We find that, prior to the intervention, households on average hold correct beliefs about their aggregate consumption, but perform only poorly in ranking energy intensities at the appliance-level. During our intervention, households' appliance-level beliefs become more accurate in the Disaggregate Feedback group, but not in the Aggregate Feedback group.

Sub-Treatment Arms on Types of Appliance-Level Feedback. We implement four sub-treatment arms to test whether additional monetary incentives for reaching appliance-specific savings targets, appliance-specific social comparisons, or a combination of both increases the conservation effect of disaggregate feedback. As shown in Subsection A5.5, we do not find that this is the case.

Persistence of Treatment Effects and Attrition. While statistically insignificant, the treatment effects decrease slightly over time (for details, see Subsection A5.6 and Table A12). The decline may be driven by a seasonal reduction in baseline electricity use of more than 40% between the beginning and the end of the study period (see Table 2). The decline may also be related to attrition: our ATE estimates increase slightly when we restrict the sample to a balanced panel, i.e., when we only include participants with complete data transmission during the core study period (Table A10 of the Appendix).

5 Consumer Surplus

In this section, we go beyond estimating conservation effects and quantify the impact of appliance-level feedback on consumer surplus. In Section 2, we have identified the following sufficient statistics to point-identify and bound the consumer surplus gains from disaggregate and aggregate feedback, respectively: the relative appliance-specific treatment effects of disaggregate feedback on input use, $\Delta y_j^s / y_j$ (and, equivalently, appliance-specific expenditure changes E_j^s), the relative effect of aggregate feedback on total electricity use $\Delta y^a / y$ (and, equivalently, total expenditure changes E^a), as well as the price elasticities of energy service demand for every appliance category j, η_j.

A measure of the treatment effects we have identified as sufficient statistics is directly available from our empirical results in Table 2. To estimate the price elasticities of appliance-level energy service demand η_j, we employ cross-sectional variation in our dataset that stems from the fact that similar households in terms of observable characteristics pay different electricity prices, in particular owing to transmission charges that vary strongly by region.

Our elasticity estimates, depicted in Column (4) of Table 3, show that appliance-level consumptions are particularly elastic for the category *Dryer*, where the estimate reaches -3.42. These estimates reflect that consumers can easily substitute this energy service by, for example, dry-hanging clothes. For the categories always-on, refrigeration, and washing, we obtain much smaller elasticities of -0.29 to -0.55. Our appliance-level elasticity estimates imply a household-level elasticity of -0.39, which is close to the estimate of -0.44, taken from Frondel et al. (2019).[14] This finding reduces concerns that our cross-sectional identification strategy yields strongly biased estimates.

[14]The elasticity of total consumption can be calculated as follows: $\eta = \sum_j \eta_j (y_j / y)$, where η_j and y_j denote the elasticity and the consumption level for appliance j, respectively, and $y = \sum_j y_j$ denotes total consumption.

Table 3: Changes in Consumer Surplus from Appliance-Level Feedback

a) Changes in Consumer Surplus from Disaggregate Feedback (ΔCS^s)

	(1)	(2)	(3)	(4)	(5)	(6)
	$\Delta y_j^s / y_j$	Avg. use in kWh/a	ΔE_j in EUR/a	η_j	$b_j^s / 2e_j$	ΔCS^s in EUR/a
Always-On	−0.002	885.35	−0.32	−0.31	−0.00	0.00
Refrigeration	−0.007	372.88	−0.66	−0.40	−0.01	0.01
Dish-Washer	−0.091	117.69	−2.54	−0.40	−0.11	0.29
Washing	−0.028	184.07	−1.21	−0.55	−0.03	0.03
Dryer	−0.439	69.10	−7.22	−3.42	−0.06	0.46
Oven	0.024	85.47	0.50	−1.07	0.01	0.01
Other appl.	−0.072	2,145.96	−36.88	−0.29	−0.12	4.60
Total		3,860.51	−48.34	−0.39		5.39

b) Changes in Consumer Surplus from Aggregate Feedback (ΔCS^a)

	$\Delta y^a / y$	Avg. use in kWh/a	ΔE in EUR/a	$(b^a/2 + b^s)/e_j$	ΔCS^a in EUR/a
Lower bound:	−0.009	3,860.51	−7.42	−0.06	0.44
Upper bound:	−0.009	3,860.51	−7.42	−0.44	3.29

Notes for Panel a): $\Delta y^a / y$ and $\Delta y_j^s / y_j$ correspond to the point estimate for group A from Table 2 Panel a), Column 2, and Panel b), Column 2, respectively. The change in Expenditures, ΔE_j^s and ΔE^a, is calculated as the product of these point estimates (Column 1), the average annual electricity use in the Aggregated Feedback (A) group (Column 2), and the average electricity price in our sample (0.238 EUR per kWh). η_j denotes the price elasticity of energy service demand with respect to the electricity price, which we estimate as described in Section 5. b_j^s / e_j denotes the relative service-specific bias, which we calculate as $-(\Delta y_j^s / 2y_j)/\eta_j$ (for derivations, see Appendix A2.1). Changes in consumer surplus are calculated as described in Equation (2), respectively.
Notes for Panel b): The bounds for changes in consumer surplus are calculated as described in Equation (3).

Based on the appliance-level elasticities and ATEs, we estimate how consumer surplus responds to the provision of disaggregate feedback. Column (1) of Table 3 reproduces the appliance-level ATEs in response to disaggregate feedback from Table 2b, which correspond to $\Delta y_j^s / y_j$ in our model. We then estimate the relative bias as $b_j^s / e_j = -(\Delta y_j^s / 2y_j)/\eta_j$, which is depicted in Column (5). We find that relative biases are negative for all appliance categories, except for *Oven*. Negative biases are most pronounced for the categories *Dish-Washer* and *Other Appliances*, where consumers underestimate energy intensities by 11% and 12%, respectively. For the categories *Washing* and *Dryer*, we obtain less pronounced biases of −3% to −6%. The relatively low bias estimate for driers illuminates that a large behavioural response is not necessarily indicative of a large misperception, but may also be caused by a high price elasticity. This observation underlines the general point that welfare effects of feedback cannot be derived from changes in consumption alone. The bias for the categories *Refrigeration*, *Oven*, and *Always-On* is close to zero.

To determine how disaggregate feedback changes consumer surplus, we calculate the appliance-level change in expenditures, ΔE_j^s, as the product of the relative ATE, $\Delta y_j^s / y_j$ (Column 1 of Table 3), the average consumption level, y_j (Column 2), and the electricity price p. We then multiply the change in expenditures with the relative bias ($b_j^s / 2e_j$) to obtain the change in consumer surplus that can be attributed to every appliance category (Column 6). Summing over all categories, we find that total consumer surplus increases by 5.4 EUR per annum and household (last row of Column 6), which is substantially less than the 48.3 EUR decrease in annual expenditures (last row of Column 3).

There are two main reasons why changes in total expenditures are an incorrect measure for changes in consumer surplus. First, less consumption of an energy service not only reduces expenditures, but also utility. The reduction in utility is proportional to the relative bias from Column (5), which in our setting is at most 12% (in absolute value). Second, consumer surplus can also rise when more accurate beliefs lead to higher consumption of an energy service, despite the fact that expenditures increase. In our setting, this occurs for the category *Oven*, for which our estimates imply that consumers slightly overestimate energy intensity.

In Panel b), we present our bounds for the consumer surplus gains in response to aggregate feedback. Our point estimate of a 0.9% reduction in electricity use translates into a reduction of expenditures by 7.42 EUR per annum and household. Yet, as shown by the bounds for the relative aggregate bias (second-last row of Panel b), our estimates imply that only 6% to maximally 44% of these savings translate into consumer welfare gains. Hence, we bound the consumer surplus effects between 0.44 and 3.29 EUR per annum and household. Even the upper bound is lower than the consumer surplus gain we estimate for disaggregate feedback (5.4 EUR per annum and household). Providing such feedback in addition to aggregate feedback is thus crucial to reap the full potential of smart metering.

To test the sensitivity of our results to these elasticity estimates, we conduct comprehensive checks based on aggregate elasticity estimates from the literature (see, e.g., Frondel et al. 2019). The major outcome of these checks is that our aggregate consumer surplus estimates are robust to a variety of appliance-level elasticities that are consistent with aggregate elasticity estimates from the literature (see Appendix A6 for details on the estimation and Appendix Table A15 for sensitivity checks). In particular, we find that the consumer surplus gain from disaggregate feedback lies between 4.8 and 7.9 EUR per household and year, while the minimum and maximum gain from aggregate feedback ranges from 0.3 to 0.6 EUR and 3.1 to 7.5 EUR, respectively. Our checks also reveal that the appliance-level contributions are relatively sensitive to the specific elasticities use. For example, while Table 3 suggests that the main consumer surplus gains arise in the category *Other Appliances*, this is no longer the case under alternative assumptions regarding appliance-level elasticities.

Overall, our findings demonstrate that official cost-benefit analyses in the EU and the U.S., for example, overestimate the consumer surplus gains from feedback substantially. Our results have shown that only 11% (5.39 EUR / 48.3 EUR) of the estimated expenditure savings from disaggregate feedback translate into changes in consumer surplus and only 6 to 44% of the savings realized by aggregate feedback (second last column of Table 3). Hence, traditional cost-benefit analyses overestimate consumer benefits from aggregate feedback by factor of two to twenty and the benefits from disaggregate feedback by a factor of around ten.

6 Conclusion

In this paper, we conduct a randomized controlled trial to investigate the effects of providing households with appliance-specific feedback. Our findings show that appliance-specific feedback leads to an additional electricity conservation effect of 5% beyond the savings induced by aggregate feedback alone. Hence, the provision of appliance-level feedback should be an integral part of the smart meter roll-out in the EU and beyond. Our evidence implies that the high effectiveness of appliance-level feedback stems from its ability to overcome appliance-specific misperceptions of energy intensities, in line with previous evidence (Attari et al., 2010). Correcting these misperceptions via feedback allows consumers to more efficiently use their appliances and to reduce total electricity use and cost.

We also provide a novel tool for evaluating the consume surplus gains from feedback based on few sufficient statistics. As we show, such gains can be calculated as the *weighted* sum of appliance-level energy cost savings. The weights are given by consumers' relative biases, which measure consumers' misperception of input intensities. This contrasts with the approach pursued in current cost-benefit analyses (Giordano et al., 2012; Faruqui et al., 2011) that equalizes consumer surplus gains by the expenditure savings. As the correct weights typically add up to less than one, current cost-benefit analyses tend to overestimate the gains in consumer surplus substantially. We also derive how the weights can be identified from observable household behaviours, in particular the behavioural response to feedback and the appliance-level responsiveness to electricity price changes.

The relevance of our findings extends beyond the context of smart meter feedback. Consumers hold misperceptions not only about the electricity consumption of appliances, but also about the effectiveness of fitness activities, the caloric content of foods, and benefits from schooling returns, for example (Attari et al. 2010, Bollinger et al. 2011, Jensen 2010). In such settings, eliminating biases via feedback holds the promise to improve the effectiveness of physical exercise and the nutritional quality of diets. Our paper provides policy makers with a tool for weighting the cost of feedback interventions against their benefits. Such tools are particularly important as advances in digitalization will likely raise the policy relevance of various forms of feedback in the future.

Appendix (For Online Publication)

Supplementary Figures

Supplementary Tables

A1 Behavioural Predictions of Feedback on Input Use

Let consumers have the following quasi-linear utility function:

$$U(\mathbf{x}, z) = u(\mathbf{x}) + z, \tag{6}$$

where $u(\mathbf{x})$ denotes the utility from consuming J energy services denoted by the vector $\mathbf{x} = (x_1, \ldots, x_J)'$, and z denotes the numeraire good, whose price is normalized to 1. The consumption of energy service j requires energy inputs of $y_j = x_j e_j$, where e_j denotes the energy intensity of energy service x_j. Consumers maximize their utility subject to the budget constraint $w = z + \sum_j y_j p$, where w denotes their exogenous income and p denotes the price of energy. Let consumers have biased perceptions of energy intensities $\tilde{e}_j = e_j + b^a + b^s_j$, where b^a and b^s_j denote the aggregate and appliance-specific bias, respectively.

A1.1 Feedback Removing the Aggregate Bias

The first-order condition of consumers' utility maximization with respect to service demand x_i is given by:

$$u_{x_i} = (e_i + b^a + b^s_i)p \tag{7}$$

Let $H = (u_{x_i x_j})$ denote the Hessian matrix. Letting h_{ij} denote the entry on the ith row and jth column of H^{-1}, we have the following lemma:

Lemma 1. *For all $i = 1, \ldots, n$,*

$$\frac{\partial x_i}{\partial p} = \sum_{j=1}^{n} h_{ij} e_j.$$

Proof. We first determine the comparative statics of the first-order condition with respect to p:

$$\sum_{j=1}^{n} u_{x_i x_j} \frac{\partial x_j}{\partial p} = e_i + b^a + b^s_i.$$

Taking $b^a + b^s_i = 0$ and then taking inverse, we get:

$$H \begin{pmatrix} \frac{\partial x_1}{\partial p} \\ \vdots \\ \frac{\partial x_n}{\partial p} \end{pmatrix} = \begin{pmatrix} e_1 \\ \vdots \\ e_n \end{pmatrix} \implies \begin{pmatrix} \frac{\partial x_1}{\partial p} \\ \vdots \\ \frac{\partial x_n}{\partial p} \end{pmatrix} = H^{-1} \begin{pmatrix} e_1 \\ \vdots \\ e_n \end{pmatrix},$$

completing the proof. □

We consider the impact of a full removal of aggregate bias, i.e. $\Delta b^a = -b^a$. Let $\Delta x^a_i = \frac{\partial x_i}{\partial b^a} \Delta b^a$ denotes the change in x_i from such a change. We claim the following:

Proposition 1. *The overall change in aggregate input use as a result of removal of aggregate biases is given by*

$$\sum_{i=1}^{n} e_i \Delta x_i^a = -b^a p \sum_{i=1}^{n} \frac{\partial x_i}{\partial p}.$$

Proof. To evaluate the impact of a change in the aggregate input use, we determine the comparative statics of this first-order condition with respect to b^a which is:

$$\sum_{j=1}^{n} u_{x_i x_j} \frac{\partial x_i}{\partial b^a} = p.$$

Using $\Delta x_i^a = \frac{\partial x_i}{\partial b^a} \Delta b^a = -b^a \frac{\partial x_i}{\partial b^a}$, we get

$$\begin{aligned}
\sum_{j=1}^{n} u_{x_i x_j} \Delta x_i^a \frac{1}{-b^a} &= p \\
\implies H \Delta \mathbf{x}^a &= -b^a \mathbf{p} \\
\implies \frac{\Delta \mathbf{x}}{-b^a} &= H^{-1} \mathbf{p},
\end{aligned}$$

where $\Delta \mathbf{x} = (\Delta x_1^a, \ldots, \Delta x_n^a)'$ and $\mathbf{p} = (p, \ldots, p)'$. Multiplying by e_i gives:

$$e_i \frac{\Delta x_i^a}{-b^a p} = \sum_{j=1}^{n} h_{ij} e_i.$$

Summing over all i yields:

$$\begin{aligned}
\sum_{i=1}^{n} e_i \frac{\Delta x_i^a}{-b^a p} &= \sum_{i=1}^{n} \sum_{j=1}^{n} h_{ij} e_i \\
&= \sum_{j=1}^{n} \sum_{i=1}^{n} h_{ji} e_i \\
&= \sum_{i=1}^{n} \frac{\partial x_i}{\partial p},
\end{aligned}$$

where the last equality stems from Lemma 1. □

Hence, we can identify the aggregate bias by :

$$\begin{aligned}
b^a &= \frac{-\sum_{i=1}^{n} e_i \Delta x_i^a}{p \sum_{j=1}^{n} \frac{\partial x_j}{\partial p}} \\
&= \frac{-\Delta y^a}{p \sum_{j=1}^{n} \frac{\partial x_j}{\partial p}},
\end{aligned}$$

where $\Delta y^a = \sum_i^n \Delta y_i^a = \sum_i^n e_i \Delta x_i^a$ denotes the total change in input use. Hence, as long as standard assumptions on the price response of service demand x_j hold ($\partial x_j / \partial p < 0$), the sign of the aggregate bias can be directly inferred from the sign of the overall demand change. Specifically, a reduction of total input use implies a negative bias, i.e., an aggregate underestimation of input use intensities.

A1.2 Feedback Removing Disaggregate Biases

To evaluate how a removal of disaggregate biases effects total input use, we again start from the first-order condition of consumers' utility maximization with respect to service demand x_j:

$$u_{x_j} = (e_j + b_j^s)p$$

The aggregate bias is set to zero as we have investigated this case in the previous subsection.

In the absence of any structure on disaggregate biases, the effect of removing disaggregate biases is indeterminate (the case of a removal of an aggregate bias extends to the the cases when all biases are positive or negative).

Hence, we put some structure on belief biases that ensures that their disaggregate biases are not naive in the sense of implying an aggregate bias. To that end, we let input intensity beliefs $\tilde{e}_j = \alpha e_j + (1-\alpha)e$, where $\alpha \in [0,1)$ is the weight that individuals put on the correct intensity and e denotes a belief that is consistent with a correct perception of total input use, but does not distinguish between different appliances, thus satisfying $\sum x_j e_j = (\sum x_j)e$ (see Gabaix 2017; Enke and Graeber 2023). Beyond ensuring that appliance-level biases do not imply an aggregate misperception, this specification also implies that eliminating appliance-level feedback may increase input intensity beliefs for some appliances and reduce them for others. Hence, the first order condition of consumers' maximization is:

$$u_{x_i} = (\alpha e_i + (1-\alpha)e)p.$$

Let $\Delta x_i^s = (1-\alpha)\frac{\partial x_i}{\partial \alpha}$ denote the change in in x_i from the removal of appliance-specific bias. Then we get the following result:

Proposition 2. *The overall change in aggregate input use as a result of removal of appliance-specific biases is given by*

$$\sum_{i=1}^{n} e_i \Delta x_i^s = p(1-\alpha)\sum_{i=1}^{n}(e_i - e)\frac{\partial x_i}{\partial p}. \tag{8}$$

Furthermore, assuming that $e_1 \geq e_2 \ldots \geq e_n$ (with at least one strict inequality), aggregate input use decreases if

$$\eta_1 \leq \eta_2 \leq \ldots \leq \eta_n$$

with at least one strict inequality, where $\eta_i = (\partial x_i/\partial p)(p/x) = (\partial y_i/\partial p)(p/y)$ denotes the elasticity of input use with respect to the input price p.

Proof. To evaluate the impact of removing the appliance-specific bias, we determine the comparative statics of this first-order condition with respect to α and p. After rearranging, we obtain the following system of equations (for $i = 1, \ldots, n$):

$$\sum_{j=1}^{n} e_j h_{ij} = \frac{\partial x_i}{\partial p} \tag{9}$$

$$\sum_{j=1}^{n} e h_{ij} + \frac{\Delta x_i^s}{p} \tilde{\alpha} = \frac{\partial x_i}{\partial p}, \tag{10}$$

where $\tilde{\alpha} = (1-\alpha)^{-1}$ and h_{ij} is the (i,j)-th entry of the inverse of the Hessian matrix of u.

For each $i = 1, \ldots, n$ we can rewrite (10) as

$$\sum_{j=1}^{n} e_i h_{ij} + \frac{e_i \Delta x_i^s}{ep} \tilde{\alpha} = \frac{e_i}{e} \frac{\partial x_i}{\partial p}. \tag{11}$$

Subtracting (11) from (9), we get for $i = 1, \ldots, n$:

$$\sum_{j \neq i} (e_j - e_i) h_{ij} - \frac{e_i \Delta x_i^s}{ep} \tilde{\alpha} = \frac{e - e_i}{e} \frac{\partial x_i}{\partial p}. \tag{12}$$

Summing (12) for all i, we get

$$-\frac{\sum_{i=1}^{n} e_i \Delta x_i^s}{ep} \tilde{\alpha} = \frac{1}{e} \sum_{i=1}^{n} (e - e_i) \frac{\partial x_i}{\partial p}.$$

Rearranging gives us (8) as desired.

We now explore the conditions under which total input use reduces, which occurs if:

$$\sum_{i=1}^{n} e_i \Delta x_i^s = p(1-\alpha) \sum_{i=1}^{n} (e_i - e) \frac{\partial x_i}{\partial p} < 0. \tag{13}$$

Using that $p > 0$, $1 - \alpha > 0$, this reduces to finding conditions such that

$$\sum_{i=1}^{n} (e_i - e) \frac{\partial x_i}{\partial p} < 0. \tag{14}$$

Using the definition of $e = \sum_{i=1}^{n} g_i e_i$ as sum of e_i (with weights $g_i = x_i / \sum_j x_j$) we obtain the the inequality:

$$\sum_{i=1}^{n} \frac{\partial x_i}{\partial p} \sum_{j=1}^{n} g_j (e_i - e_j) < 0. \tag{15}$$

Multiplying by $\sum_j x_j$ and rearranging, we can rewrite this inequality to get:

$$\sum_{i,j:i<j} (e_i - e_j) \left[x_j \frac{\partial x_i}{\partial p} - x_i \frac{\partial x_j}{\partial p} \right] < 0 \tag{16}$$

Without loss of generality, assume that $e_1 \geq e_2 \ldots \geq e_n$ (with at least one strict inequality). Then, a sufficient condition for a reduction in total input use is given by:

$$\frac{\partial x_1}{\partial p} x_1^{-1} \leq \frac{\partial x_2}{\partial p} x_2^{-1} \leq \ldots \leq \frac{\partial x_n}{\partial p} x_n^{-1},$$

with at least one strict inequality. Multiplying by p yields:

$$\eta_1 \leq \eta_2 \leq \ldots \leq \eta_n,$$

as was to be shown. □

One possible functional form that satisfies the condition in Proposition 2 would be that for all $i = 1, \ldots, n$,

$$\frac{\partial x_i}{\partial p} = -C x_i e_i,$$

where $C > 0$ is a constant. We have

$$\sum_{i,j:i<j} (e_i - e_j) \left[x_j \frac{\partial x_i}{\partial p} - x_i \frac{\partial x_j}{\partial p} \right] = \sum_{i,j:i<j} -C x_i x_j (e_i - e_j)^2.$$

We will thus get following expression for change in total input use:

$$\sum_{i=1}^{n} e_i \Delta x_i^s = -\frac{Cp(1-\alpha)}{\sum_{i=1}^{n} x_i} \sum_{i,j:i<j} x_i x_j (e_i - e_j)^2,$$

which is clearly negative.

One could weaken the condition to

$$\frac{\partial x_i}{\partial p} = -C_i x_i e_i,$$

provided that the sequence $C_1, \ldots, C_n > 0$ satisfies the property $C_i / C_j \geq e_j / e_i$ whenever $e_i \geq e_j$, with inequality strict when $e_i > e_j$. In this case, we have

$$\sum_{i,j:i<j} (e_i - e_j) \left[x_j \frac{\partial x_i}{\partial p} - x_i \frac{\partial x_j}{\partial p} \right] = \sum_{i,j:i<j} -x_i x_j (C_i e_i - C_j e_j)(e_i - e_j),$$

hence the term is indeed negative since the terms $(C_i e_i - C_j e_j)$ and $(e_i - e_j)$ share the same sign.

Lemma 2. *The bias parameter $1 - \alpha$ can be identified empirically as*

$$1 - \alpha = \frac{\sum_{i=1}^{n} e_i \Delta x_i^s}{p \sum_{i=1}^{n} (e_i - e) \frac{\partial x_i}{\partial p}} \tag{17}$$

Proof. This follows immediately from rearranging equation (13). □

The expression in equation (17) has a straightforward interpretation: the numerator in the equation is the change in energy use due to the removal of the error $(1 - \alpha)(e_i - e)$ in the perception of service intensities for appliance i. The denominator expresses by how much energy use would change due to a change in perceptions $\Delta\alpha = 1$: it would act like a price change of $(e_i - e)p$ for service i, and must hence be weighted with the slope of the demand $\frac{\partial x_i}{\partial p}$ for service i.

A1.3 Belief Updating from Information for a Subset of Appliances

In this section, we explore what a consumer can learn from disaggregate consumption feedback for a group of appliances with respect to the consumption of another group of appliances where no feedback is given.

To formalize that problem, let y_1 denote the consumption of a group of appliances for which a consumer obtains aggregate feedback, and y_2 the consumption of a group of appliances for which a consumer obtains no disaggregate feedback. We also assume that the consumer observes aggregate consumption $Y = y_1 + y_2$. A consumer has prior beliefs about appliance 1 and 2 that are normally distributed: $y_i \sim N(\mu_y, \sigma_y^2)$ for $i = 1, 2$, where μ_y and σ_y denote the mean and variance of all types of appliances that exist. We also assume that a consumer obtains a normally distributed signal regarding the consumption of both appliances that consists of "truth plus noise": $s_i = y_i + u_i$ where $u_i \sim N(0, \sigma_{u_i}^2)$ denotes the noise that is assumed to be independent of true consumption levels y_i. The precision of the signal i is thus given by $1/u_i$. Finally, we assume that the consumer processes signals via Bayesian updating.

This model setup captures the core features of our experimental setting. Consumers may be well aware about the distribution of energy uses by different appliances, but they are fundamentally unsure about which appliance uses more (or less) energy than another (Attari et al., 2010). Note that, for simplicity, we cast the problem about a problem of inferring energy uses rather than energy intensities. Both problems are equivalent as consumers can easily observe the appliance usage x, but not the energy intensity e (and, thus, energy use $y = x \cdot e$).

We now explore how obtaining more precise information about the consumption for certain appliances affects the beliefs concerning the appliances where no additional information is received. In particular, we model better information as an increase in the precision of the signal obtained for the second group of appliances, for example because disaggregate feedback is provided. Formally, we assess how the conditional distribution of y_1 given s_1, s_2, and Y changes as σ_{u_2} decreases (the precision of the signal $1/\sigma_{u_2}$ increases).[15]

Before deriving the conditional distribution, we first make some definitions. Let μ_x denote the expected value of a random variable X. Furthermore, let $\mathbf{Z} = [y_1, s_1, s_2, Y]^T$ denote the vector of the variables of interest in this setting. The vector of signals a consumer obtains is denoted by $\mathbf{S} = [s_1, s_2, Y]^T$. The covariance matrix Σ of $\mathbf{Z}$ is then given by:

$$\Sigma = \begin{bmatrix} \sigma_y^2 & \sigma_y^2 & 0 & \sigma_y^2 \\ \sigma_y^2 & \sigma_y^2 + \sigma_{u_1}^2 & 0 & \sigma_y^2 \\ 0 & 0 & \sigma_y^2 + \sigma_{u_2}^2 & \sigma_y^2 \\ \sigma_y^2 & \sigma_y^2 & \sigma_y^2 & \sigma_y^2 + \sigma_y^2 \end{bmatrix}$$

[15] In this section, we used ChatGPT to reproduce textbook results on conditioning results for normal distributions. All subsequent derivations were produced by us.

We can partition

$$\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{14} \\ \Sigma_{41} & \Sigma_{44} \end{bmatrix},$$

where the sub-matrix Σ_{44} is the covariance matrix for the variables s_1, s_2, and Y:

$$\Sigma_{44} = \begin{bmatrix} \sigma_y^2 + \sigma_{u_1}^2 & 0 & \sigma_y^2 \\ 0 & \sigma_y^2 + \sigma_{u_2}^2 & \sigma_y^2 \\ \sigma_y^2 & \sigma_y^2 & \sigma_y^2 + \sigma_y^2 \end{bmatrix}$$

General formula for conditional distribution

The conditional distribution of y_1 given a vector of realized signals s is given by

$$y_1|s \sim N(\mu_{y_1|s}, \Sigma_{y_1|s}),$$

where

$$\mu_{y_1|s} = \mu_y + \Sigma_{14}\Sigma_{44}^{-1} \left(s - \begin{bmatrix} \mu_y \\ \mu_y \\ \mu_y + \mu_y \end{bmatrix} \right) \tag{18}$$

and

$$\Sigma_{y_1|s} = \sigma_y^2 - \Sigma_{14}\Sigma_{44}^{-1}\Sigma_{41} \tag{19}$$

denote the conditional mean and variance of y_1, which capture the posterior beliefs of participants after having observed signals about appliance 1, 2, and total energy use.

Equations (18) and (19) capture the rationale how a consumer employing Bayes' rule incorporates information about signals s_1, s_2, and Y into her beliefs. Equation (18) shows how the consumer updates her posterior mean belief regarding the consumption of appliance 1. It shows that updating is strong when a) the signals received z deviate from their expected value, and b) when the term $\Sigma_{14}\Sigma_{44}^{-1}$ is large, which captures the informativeness of a given signal. Equation (19) describes how the consumer updates her posterior variance regarding the consumption of appliance 1. It shows that the posterior second moment consists of the variance of y_1 less the part of the variance that is explained by the signals z, $\Sigma_{14}\Sigma_{44}^{-1}\Sigma_{41}$.

In the following, we explore how a higher precision of a signal for good 2 affects the inferences a consumer makes from that signal about the consumption of good 1. This will allow us to explore in what way consumers rationally adjust their beliefs for a group of appliances when disaggregate consumption feedback is provided for another group of appliances.

Derivation of conditional distribution in our setting

We now derive the expressions $\Sigma_{14}\Sigma_{44}^{-1}$ and $\Sigma_{14}\Sigma_{44}^{-1}\Sigma_{41}$ in our setting. For that purpose, we partition the sub-matrix Σ_{44} into the following components:

$$A = \begin{bmatrix} \sigma_y^2 + \sigma_{u_1}^2 & 0 \\ 0 & \sigma_y^2 + \sigma_{u_2}^2 \end{bmatrix}, \quad B = \begin{bmatrix} \sigma_y^2 \\ \sigma_y^2 \end{bmatrix}, \quad C = B^T = \begin{bmatrix} \sigma_y^2 & \sigma_y^2 \end{bmatrix}, \quad D = \sigma_y^2 + \sigma_y^2$$

The Schur complement S is given by

$$S = D - CA^{-1}B = \sigma_y^2 + \sigma_y^2 - \frac{\sigma_y^4}{\sigma_y^2 + \sigma_{u_1}^2} - \frac{\sigma_y^4}{\sigma_y^2 + \sigma_{u_2}^2}$$

Let the "signal-to-noise" ratio for appliance 1 be denoted by by:

$$\alpha = \frac{\sigma_y^2}{\sigma_y^2 + \sigma_{u_1}^2}$$

and for appliance 2 by:

$$\beta = \frac{\sigma_y^2}{\sigma_y^2 + \sigma_{u_2}^2}.$$

We can rewrite S as follows:

$$S = \sigma_y^2(1-\alpha) + \sigma_y^2(1-\beta). \tag{20}$$

Note that a higher precision of signal 2 (lower $\sigma_{u_2}^2$) increases β and decreases S (and thus increases S^{-1}).

We can compute the inverse Σ_{44}^{-1} as follows:

$$\Sigma_{44}^{-1} = \begin{bmatrix} A^{-1} + A^{-1}BS^{-1}CA^{-1} & -A^{-1}BS^{-1} \\ -S^{-1}CA^{-1} & S^{-1} \end{bmatrix} = \begin{bmatrix} A^{-1} & 0 \\ 0 & 0 \end{bmatrix} + S^{-1} \begin{bmatrix} A^{-1}BCA^{-1} & -A^{-1}B \\ -CA^{-1} & 1 \end{bmatrix}$$

Note that $A^{-1}B = \begin{bmatrix} \alpha \\ \beta \end{bmatrix}$ and $CA^{-1} = \begin{bmatrix} \alpha & \beta. \end{bmatrix}$
We thus have

$$\begin{bmatrix} A^{-1}BCA^{-1} & -A^{-1}B \\ -CA^{-1} & 1 \end{bmatrix} = \begin{bmatrix} \alpha^2 & \alpha\beta & -\alpha \\ \alpha\beta & \beta^2 & -\beta \\ -\alpha & -\beta & 1 \end{bmatrix}.$$

We are interested in the quantity $\Sigma_{14}\Sigma_{44}^{-1}\Sigma_{41}$, where

$$\Sigma_{14} = \begin{bmatrix} \sigma_y^2 & 0 & \sigma_y^2 \end{bmatrix} = \sigma_y^2 \begin{bmatrix} 1 & 0 & 1 \end{bmatrix},$$

and

$$\Sigma_{41} = \Sigma_{14}^T = \sigma_y^2 \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix}.$$

We have

$$\begin{bmatrix} 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} A^{-1} & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix} = \frac{1}{\sigma_y^2 + \sigma_{u_1}^2}.$$

We also have

$$\begin{bmatrix}1 & 0 & 1\end{bmatrix}\begin{bmatrix}\alpha^2 & \alpha\beta & -\alpha\\ \alpha\beta & \beta^2 & -\beta\\ -\alpha & -\beta & 1\end{bmatrix}\begin{bmatrix}1\\0\\1\end{bmatrix} = \alpha^2 - 2\alpha + 1 = (\alpha-1)^2 = \left(\frac{\sigma_{u_1}^2}{\sigma_y^2+\sigma_{u_1}^2}\right)^2.$$

It follows that

$$\Sigma_{14}\Sigma_{44}^{-1}\Sigma_{41} = \frac{\sigma_y^4}{\sigma_y^2+\sigma_{u_1}^2}\left[1 + S^{-1}\frac{\sigma_{u_1}^4}{\sigma_y^2+\sigma_{u_1}^2}\right]. \tag{21}$$

Furthermore, we have that:

$$\begin{aligned}\Sigma_{14}\Sigma_{44}^{-1} &= \sigma_y^2\begin{bmatrix}1 & 0 & 1\end{bmatrix}\{\begin{bmatrix}\frac{1}{\sigma_y^2+\sigma_{u_1}^2} & 0 & 0\\ 0 & \frac{1}{\sigma_y^2+\sigma_{u_2}^2} & 0\\ 0 & 0 & 0\end{bmatrix} + S^{-1}\begin{bmatrix}\alpha^2 & \alpha\beta & -\alpha\\ \alpha\beta & \beta^2 & -\beta\\ -\alpha & -\beta & 1\end{bmatrix}\}\\ &= \sigma_y^2\begin{bmatrix}1 & 0 & 1\end{bmatrix}\begin{bmatrix}\frac{1}{\sigma_y^2+\sigma_{u_1}^2} & 0 & 0\\ 0 & \frac{1}{\sigma_y^2+\sigma_{u_2}^2} & 0\\ 0 & 0 & 0\end{bmatrix} + \sigma_y^2\begin{bmatrix}1 & 0 & 1\end{bmatrix}S^{-1}\begin{bmatrix}\alpha^2 & \alpha\beta & -\alpha\\ \alpha\beta & \beta^2 & -\beta\\ -\alpha & -\beta & 1\end{bmatrix}\\ &= \sigma_y^2\begin{bmatrix}\frac{1}{\sigma_y^2+\sigma_{u_1}^2} & 0 & 0\end{bmatrix} + \sigma_y^2 S^{-1}\begin{bmatrix}\alpha^2-\alpha & \alpha\beta-\beta & 1-\alpha\end{bmatrix}\end{aligned}$$

Rewriting thus gives:

$$\Sigma_{14}\Sigma_{44}^{-1} = \sigma_y^2 S^{-1}\begin{bmatrix}\frac{S}{\sigma_y^2+\sigma_{u_1}^2}+\alpha^2-\alpha & \beta(\alpha-1) & 1-\alpha.\end{bmatrix} \tag{22}$$

Result 1: Belief updating regarding appliance 1 consumption is stronger when the signal for appliance 2 becomes more precise.

Recall that our thought experiment is that $\sigma_{u_2}^2$ decrease (e.g., in response to disaggregate feedback) and then to explore the impact of this on the precision of beliefs for y_1, i.e., the consumption of appliances not covered by disaggregate feedback. We first explore how posterior means ($\mu_{y_1|s}$) are affected.

Putting Equations (18) and (22) together, we get:

$$\mu_{y_1|s} = \mu_y + \underbrace{\sigma_y^2 S^{-1}\begin{bmatrix}\frac{S}{\sigma_y^2+\sigma_{u_1}^2}+\alpha^2-\alpha & \beta(\alpha-1) & 1-\alpha\end{bmatrix}}_{(1)}\underbrace{\left(s - \begin{bmatrix}\mu_y\\ \mu_y\\ \mu_y+\mu_y\end{bmatrix}\right)}_{(2)}. \tag{23}$$

Notice that the impact of on the posterior mean, captured in Equation (23), consists of two components. First, the strength of updating (1), and, second, the degree to which the signal obtained by a consumer deviates from his posterior mean (2).

For the purpose of our thought experiment, where σ_{u_2} decreases, we now want to assess how such a change affects the posterior mean for the consumption of appliance 1. This follows from the second element from Equation (23), which is given by:

$$\mu_y + \underbrace{\sigma_y^2 S^{-1}(\sigma_{u_2})\beta(\alpha - 1)}_{(1)} \underbrace{(s_2 - \mu_y)}_{(2)}. \tag{24}$$

Intuitively, the posterior mean for the consumption of appliance 1 consists of the prior mean μ_y, plus an updating component. The updating component consists of two terms. The first term, (1), captures the degree to which a higher signal implies an adjustment of the posterior mean. In our setting, it is always negative (as $\alpha < 1$). The second term, (2), captures what a consumer learns about appliance 2. If the signal for appliance 2 is larger than anticipated, (2) is positive and a consumer adjusts his posterior mean for appliance 1 downwards. This is because the consumer now attributes more of the aggregate consumption level to appliance 1 (and less to appliance 2).

What happens to such updating when σ_{u_2} decreases? As before, S^{-1} increases when σ_{u_2} decreases. In addition, $\beta = \sigma_y^2/(\sigma_y^2 + \sigma_{u_2}^2)$ increases when $\sigma_{u_2}^2$ decreases. As $0 \leq \alpha < 1$, this implies that (1) decreases when σ_{u_2} decreases. The intuition is as follows: Better information for appliance 2 implies that a consumer puts more weight on the signal obtained for that appliance.

In the context of our experiment, highly precise feedback for appliances 2 (through disaggregation) will imply that consumers learn more about the consumption of appliance 2 and appliance 1. If the consumer learns from signal 2 that the consumption of appliance 2 is lower than expected, we have that $s_2 - \mu_y < 0$. In our experiment this is likely the case as the consumption share of large appliances covered by disaggregation is plausibly smaller than what consumers may have suspected. Consumers will then revise their posterior mean for appliance 1 upwards, even though they have not received any specific information on these appliances.

Result 2: Beliefs about appliance 1 consumption become more precise as the signal for appliance 2 becomes more precise

Next, we discuss a second channel that explains why consumers respond to better information for disaggregate appliances by changing the consumption of other appliances. This channel is more subtle and works through the precision of signals. As above, continue to assume that electricity use of disaggregate appliances is smaller than μ_y, and that consumption of the appliances without specific feedback is above μ_y.

Because the signal for a subset of appliances becomes more precise, more households will receive a signal that $(s_2 - \mu_y) < 0$. By the same token, equation (24) is thus pushed up for these households, thus leading a larger fraction of households to (correctly) believe that $y_1 > \mu_y$, and reduce consumption of these appliances, as we observe in the data.

Formally, these forces reduce the remaining variance in the forecast of y_1, as can be seen by putting (19) and (21):

$$\Sigma_{y_1|s} = \sigma_y^2 - \Sigma_{14}\Sigma_{44}^{-1}\Sigma_{41}$$
$$= \sigma_y^2 - \frac{\sigma_y^4}{\sigma_y^2 + \sigma_{u_1}^2}\left[1 + S^{-1}(\sigma_{u_2})\frac{\sigma_{u_1}^4}{\sigma_y^2 + \sigma_{u_1}^2}\right],$$

where $S(\sigma_{u_2}) = \sigma_y^2 + \sigma_y^2 - \frac{\sigma_y^4}{\sigma_y^2+\sigma_{u_1}^2} - \frac{\sigma_y^4}{\sigma_y^2+\sigma_{u_2}^2}$.

Note that $\Sigma_{y_1|z}$ depends on σ_{u_2} only via $S^{-1}(\sigma_{u_2})$. S^{-1} increases when σ_{u_2} decreases. Hence, $\Sigma_{14}\Sigma_{44}^{-1}\Sigma_{41}$ increases as $\sigma_{u_2}^2$ decreases. As this term enters negatively, the conditional variance of y_1 decreases as σ_{u_2} decreases, which is what we wanted to show. In other words, posterior beliefs regarding the consumption of good 1 become more precise as the precision of the signal for the other group of appliances increases, for example in response to the provision of disaggregate consumption feedback.

A2 Welfare and Sufficient Statistics

A2.1 Sufficient Statistics for Evaluating Consumer Surplus

Let consumers have the following quasi-linear utility function:

$$U(\mathbf{x}, z) = u(\mathbf{x}) + z, \tag{25}$$

where $u(\mathbf{x})$ denotes the utility from consuming J energy services denoted by the vector $\mathbf{x} = (x_1, \ldots, x_J)'$, and z denotes the numeraire good, whose price is normalized to 1. The consumption of energy service j requires energy inputs of $y_j = x_j e_j$, where e_j denotes the energy intensity of energy service x_j. Consumers maximize their utility subject to the budget constraint $w = z + \sum_j y_j p$, where w denotes their exogenous income and p denotes the price of energy. Let consumers have biased perceptions of energy intensities $\tilde{e}_j = e_j + b^a + b_j^s$, where b^a and b_j^s denote the aggregate and appliance-specific bias, respectively.

We write decision utility as:

$$U^s = u(\mathbf{x}) - \sum_j p(e_j + b^a + b_j^s)x_j.$$

Utility maximization yields the FOCs with respect to energy service demand x_j:

$$u_{x_j} = p\tilde{e}_j = p(e_j + b^a + b_j^s) \;\; \forall\, j \in \{1, \ldots, J\}. \tag{26}$$

Furthermore, normative utility is:

$$U^n = u(\mathbf{x}) - \sum_j p e_j x_j.$$

We are interested in the welfare effect of a change in the bias $b \to b + \Delta b$. Its second-order approximation is given by:

$$\Delta CS = U^n(b + \Delta b) - U^n(b) = U_b^{n'} \Delta b + \frac{1}{2} \Delta b' U_{bb}^n \Delta b,$$

where U_b^n and U_{bb}^n denote the vector of first and second derivatives of normative utility with respect to the bias vector b, defined by:

$$U_b^n = \begin{bmatrix} \frac{\partial U^n}{\partial b_1} \\ \vdots \\ \frac{\partial U^n}{\partial b_k} \end{bmatrix}, \quad U_{bb}^n = \begin{bmatrix} \frac{\partial^2 U^n}{\partial b_1 \partial b_1} & \cdots & \frac{\partial^2 U^n}{\partial b_1 \partial b_k} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 U^n}{\partial b_1 \partial b_k} & \cdots & \frac{\partial^2 U^n}{\partial b_k \partial b_k} \end{bmatrix}, \quad b = \begin{bmatrix} b_1 \\ \vdots \\ b_k \end{bmatrix}, \quad \Delta b = \begin{bmatrix} \Delta b_1 \\ \vdots \\ \Delta b_k \end{bmatrix}.$$

In particular, we find that:

$$\begin{aligned} \frac{\partial U^n}{\partial b_l} &= \frac{\partial U^s}{\partial b_l} + p x_l + \sum_j p b_j \cdot \frac{\partial x_j}{\partial b_l} \\ &= \sum_j \frac{\partial u}{\partial x_j} \frac{\partial x_j}{\partial b_l} - p x_l - \sum_j p(e_j + b_j) \cdot \frac{\partial x_j}{\partial b_l} + p x_l + \sum_j p b_j \frac{\partial x_j}{\partial b_l} \\ &= \sum_j \left[\frac{\partial u}{\partial x_j} - p(e_j + b_j) \right] \frac{\partial x_j}{\partial b_l} + \sum_j p b_j \frac{\partial x_j}{\partial b_l} \\ &= \sum_j p b_j \frac{\partial x_j}{\partial b_l} \\ \frac{\partial U^n}{\partial b_l b_m} &= p \frac{\partial x_m}{\partial b_l} + \sum_j p b_j \frac{\partial^2 x_j}{\partial b_l \partial b_m} \\ &= p \frac{\partial x_m}{\partial b_l}, \end{aligned} \tag{27}$$

where, in the last step, we assume that higher-order effects of changes in bias on demand are zero.

Under additive seperability, we have that $u(x) = \sum_j u_j(x_j)$. It implies that:

$$U_b^n = \begin{bmatrix} p b_1 \frac{\partial x_1}{\partial b_1} \\ \vdots \\ p b_k \frac{\partial x_k}{\partial b_k} \end{bmatrix} \qquad U_{bb}^n = \begin{bmatrix} p \frac{\partial x_1}{\partial b_1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & p \frac{\partial x_k}{\partial b_k} \end{bmatrix}.$$

Hence, we can express the change in consumer surplus as follows:

$$\begin{aligned}
\Delta CS = U^n(b+\Delta b) - U^n(b) &= U_b^{n'}\Delta b + \frac{1}{2}\Delta b' U_{bb}^n \Delta b \\
&= \sum_j p b_j \frac{\partial x_j}{\partial b_j}\Delta b_j + \frac{1}{2}\sum_j p \Delta b_j \frac{\partial x_j}{\partial b_j}\Delta b_j \\
&= \sum_j p\left(b_j + \frac{1}{2}\Delta b_j\right)\frac{\partial x_j}{\partial b_j} \\
&= \sum_j \frac{b_j \frac{1}{2}\Delta b_j}{e_j}\Delta E_j(b,\Delta b),
\end{aligned} \tag{28}$$

where $\Delta y_j(b, \Delta b)$ denotes the change in expenditures from an initervention that induces a belief change from b to $b + \Delta b$.

Starting from a situation where both aggregate and disaggregate biases exist, we now explore how the total change in consumer surplus from a debiasing intervention, ΔCS, can be split into two parts: First, a consumer surplus change from the removal of aggregate bias, ΔCS^a and, second, a corresponding change from removing disaggregate biases, ΔCS^s. To derive the respective consumer surplus changes, we first evaluate an intervention that removes the aggregate bias b^a (e.g., through aggregate feedback) and then proceed to evaluate an intervention that additionally removes disaggregate biases (e.g., through additionally presenting appliance-specific feedback).

The effect on consumer surplus from removing aggregate biases is given by applying Equation (28), noting that $b_j = b^a + b_j^s, \Delta b_j = -b^a$, and that $\Delta E_j(b_j, \Delta b_j) = \Delta E_j^a$. This yields:

$$\begin{aligned}
\Delta CS^a &= \sum_j \frac{b^a + b_j^s - \frac{1}{2}b^a}{e_j}\Delta E_j^a \\
&= \sum_j \frac{\frac{1}{2}b^a + b_j^s}{e_j}\Delta E_j^a.
\end{aligned} \tag{29}$$

Similarly, the effect on consumer surplus from removing disaggregate biases follows from Equation (28), noting that $b_j = b_j^s, \Delta b_j = -b_j^s$, and that $\Delta E_j(b_j, \Delta b_j) = \Delta E_j^s$. Hence, we obtain:

$$\Delta CS^s = \sum_j \frac{b^s}{2e_j}\Delta E_j^s. \tag{30}$$

Derivation of sufficient statistics:

In order to implement our welfare formulas, we need to express the relative aggregate and disaggregate bias in terms of observables. To do so, we totally differentiate the first-order condition of consumer maximization (Equation 26) with respect to disaggregate biases b_j^s, the aggregate bias b^a, and the price p. These comparative statics yield the following system of equations:

$$b_j^s : H \begin{bmatrix} \frac{\partial x_1}{\partial b_j^s} \\ \vdots \\ \vdots \\ \vdots \\ \frac{\partial x_k}{\partial b_j^s} \end{bmatrix} = \begin{bmatrix} 1(j=1) \\ \vdots \\ \vdots \\ \vdots \\ 1(j=J) \end{bmatrix} \cdot p \qquad \Leftrightarrow \qquad \frac{\partial x_k}{\partial b_j^s} = H^{-1} \begin{bmatrix} 1(j=1) \\ \vdots \\ \vdots \\ \vdots \\ 1(j=J) \end{bmatrix} \tag{31}$$

$$p : H \begin{bmatrix} \frac{\partial x_1}{\partial p} \\ \vdots \\ \frac{\partial x_k}{\partial p} \end{bmatrix} = \begin{bmatrix} e_1 \\ \vdots \\ e_k \end{bmatrix} \qquad \Leftrightarrow \qquad \frac{\partial x}{\partial p} = H^{-1} e \tag{32}$$

$$b^a : H \begin{bmatrix} \frac{\partial x_1}{\partial b^a} \\ \vdots \\ \frac{\partial x_k}{\partial b^a} \end{bmatrix} = p \qquad \Leftrightarrow \qquad \frac{\partial x}{\partial b^a} = H^{-1} p, \tag{33}$$

where $1(\cdot)$ denotes the indicator function. Under additive separability, we have that:

$$H = \begin{bmatrix} u_{x_1 x_1} & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & u_{x_k x_k} \end{bmatrix} \qquad H^{-1} = \begin{bmatrix} \frac{1}{u_{x_1 x_1}} & 0 & \\ \vdots & \ddots & \vdots \\ 0 & \dots & \frac{1}{u_{x_k x_k}} \end{bmatrix}.$$

Hence, we can rewrite Equations (31), (31), and (33) as:

$$\frac{\partial y_j}{\partial b_j^s} \cdot \Delta b_j^s = \frac{\partial x_j}{\partial b_j^s} \cdot e_j \cdot \Delta b_j^s = \frac{p e_j}{u_{x_j x_j}} \cdot \Delta b_j^s = p \cdot \frac{e_j^2}{u_{x_j x_j}} \frac{\Delta b_j^s}{e_j} \tag{34}$$

$$\frac{\partial y_j}{\partial p} = \frac{\partial x_j}{\partial p} \cdot e_j = \frac{e_j^2}{u_{x_j x_j}} \tag{35}$$

$$\frac{\partial y_j}{\partial b^a} \Delta b^a = p \frac{e_j}{u_{x_j x_j}} \Delta b^a = p \frac{e_j^2}{u_{x_j x_j}} \frac{\Delta b^a}{e_j}. \tag{36}$$

Inserting Equation (35) into Equation (34), noting that $\Delta b_j^s = -b_j^s$, yields:

$$\Delta y_j^s = p \cdot \frac{\partial y_j}{\partial p} \cdot \left(\frac{-b_j^s}{e_j} \right)$$

$$\Leftrightarrow \qquad \frac{b_j^s}{e_j} = -\frac{\frac{\Delta y_j^s}{y_j}}{\eta_j}. \tag{37}$$

We then insert Equation (35) into Equation (36) to obtain:

$$\Delta y_j^a = p \cdot \frac{\partial y_j}{\partial p} \left(\frac{-b^a}{e_j} \right)$$

$$\Leftrightarrow \qquad \frac{b^a}{e_j} = -\frac{\frac{\Delta y_j^a}{y_j}}{\eta_j}. \tag{38}$$

By inserting Equation (37) into Equation (29), we obtain an expression for ΔCS^s in terms of sufficient statistics:

$$\Delta CS^s = \sum_j \frac{\frac{\Delta y_j}{y_j}}{2\eta_j} \Delta E_j^s. \tag{39}$$

To obtain bounds for the consumer welfare effect of aggregate feedback, ΔCS^a, we insert Equation (38) into Equation (30), which yields:

$$\Delta CS^a = \sum_j \left[\frac{\frac{\Delta y_j^a}{y_j}}{2\eta_j} + \frac{\frac{\Delta y_j^s}{y_j}}{\eta_j} \right] \Delta E_j^a. \tag{40}$$

In the absence of information about the change in expenditures at the appliance-level, this expression is not point identified. However, note that consumer surplus is estimated by weighting the changes in expenditures by the sum of the average aggregate and disaggregate bias (the expression in squared brackets). Hence, we derive the maximum and minimum weight that is consistent with the aggregate saving, which we observe. Multiplying these weights with the aggregate savings we observe then allows us to construct an upper and lower bound for the consumer welfare change.

More specifically, the upper bound is given by:

$$\Delta CS^a \leq p \cdot \Delta y^a \cdot \left[\frac{\frac{\Delta y^a}{y_m}}{2\eta_m} + \frac{\frac{\Delta y_m^s}{y_m}}{\eta_m} \right],$$

where m is the appliance with the **largest** bracketed term. By a similar logic, the lower bound is given by:

$$\Delta CS^a \geq p \cdot \Delta y^a \cdot \left[\frac{\frac{\Delta y^a}{y_n}}{2\eta_n} + \frac{\frac{\Delta y_n^s}{y_n}}{\eta_n} \right],$$

where n is the appliance with the **smallest** bracketed term.

A2.2 Sufficient Statistics Without Separability

In this subsection, we showcase how our approach can easily be adapted to other settings beyond electricity use. We derive the sufficient statistics for an alternative model with a general utility function $u(\mathbf{x})$, but retain appliance-specific bias of the form $\tilde{e}_i = \alpha e_i + (1 - \alpha)e$ that we use in deriving the predictions for disaggregate feedback. This subsection serves to purpose to showcase that our approach can easily be adapted to other settings beyond electricity use.

Normative utility is defined as above

$$U^n = u(\mathbf{x}) - p\sum_{i=1}^{n} e_i x_i.$$

Recall that from the consumer's perspective, the first-order condition is:

$$u_{x_i} = (\alpha(e_i - e) + e)p.$$

Hence, the derivative of normative utility with respect to the bias parameter α is:

$$\begin{aligned}\frac{\partial U^n}{\partial \alpha} &= \sum_{i=1}^{n} (u_{x_i} - e_i p)\,\frac{\partial x_i}{\partial \alpha} \\ &= p\sum_{i=1}^{n} ((\alpha(e_i - e) + e) - e_i)\,\frac{\partial x_i}{\partial \alpha} \\ &= (\alpha - 1)p\sum_{i=1}^{n}(e_i - e)\frac{\partial x_i}{\partial \alpha}.\end{aligned}$$

Assuming that $\frac{\partial^2 x_i}{\partial \alpha^2} \approx 0$, we get:

$$\frac{\partial^2 U^n}{\partial \alpha^2} \approx p\sum_{i=1}^{n}(e_i - e)\frac{\partial x_i}{\partial \alpha}.$$

Putting terms together for a second-order Taylor approximation, the increase in consumer surplus from removing biased perceptions of appliance-specific service intensities, i.e of $\Delta\alpha = 1 - \alpha$ has the following approximate effect:

$$\begin{aligned}\Delta CS &\approx \frac{\partial U^n}{\partial \alpha}(1-\alpha) + \frac{1}{2}\frac{\partial^2 U^n}{\partial \alpha^2}(1-\alpha)^2 && (41)\\ &= (\alpha - 1)p\sum_{i=1}^{n}(e_i - e)\frac{\partial x_i}{\partial \alpha} + \frac{1}{2}p\sum_{i=1}^{n}(e_i - e)\frac{\partial x_i}{\partial \alpha}(1-\alpha)^2 && (42)\\ &= (\alpha - 1)p\sum_{i=1}^{n}(e_i - e)\frac{\Delta x_i^s}{1-\alpha}(1-\alpha) + \frac{1}{2}p\sum_{i=1}^{n}(e_i - e)\frac{\Delta x_i^s}{1-\alpha}(1-\alpha)^2 && (43)\\ &= \frac{1}{2}(\alpha - 1)p\sum_{i=1}^{n}(e_i - e)\Delta x_i^s. && (44)\end{aligned}$$

In the above equations, we have made use of the assumption that $\frac{\partial^2 x_i}{\partial \alpha^2} \approx 0$. Because of this, we can express the derivative $\frac{\partial x_i}{\partial \alpha} = \frac{\Delta x_i^s}{1-\alpha}$, where Δx_i^s is the behavioural change due to the feedback intervention. This is the total change in behaviour due to $\Delta\alpha = 1 - \alpha$, as we assume that the feedback intervention removes the entire bias.

The expression in Equation (44) has the same general interpretation as in the separable case, but is more compact as the bias parameter α governs the extent of bias away from the known service intensities e_i.

The expression for ΔCS in Equation (44) still contains the parameter $1-\alpha$. It can be identified empirically using Lemma 2 (equation 17). Substituting this result, we can express the sufficient statistic for the change in consumer surplus by removing the bias $1-\alpha$ in the case of non-separable utility as

$$\Delta CS = \frac{1}{2}\frac{\sum_{i=1}^{n} e_i \Delta x_i^s}{\sum_{i=1}^{n}(e_i - e)\frac{\partial x_i}{\partial p}}\sum_{i=1}^{n}(e_i - e)\Delta x_i^s \tag{45}$$

Calculation of this sufficient statistic requires three sets of measurements:

1. The change in service (energy) use due to disaggregate feedback for all appliances i, $e_i \Delta x_i^s$.
2. The demand responses $\frac{\partial x_i}{\partial p}$ with regard to a change in the service (kWh) price p.
3. Measurements of the true service intensities e_i.

Conceptually, all these elements can be identified in an experimental setup such as ours. Notice also that it is still necessary to put some structure on the form of bias, which is achieved through our behaviourally and empirically informed formulation. Identification of the sufficient statistics in a model with non-separable utility and unspecified form of the bias would require substantially more data. In particular, it would require knowledge of the demand slopes $\frac{\partial x_i}{\partial p_i}$ that vary the price for service i, while holding all other service prices constant. In the context of electricity use, this would amount to an impossibly complicated experiment: one would need experimental treatments that change the cost of use of only one of the appliances.

A2.3 General Welfare Effects of Feedback

In this section, we prove that a de-biasing intervention increases welfare under very general assumptions. Throughout, we make the assumption that higher-order effects of bias on demand is zero, i.e. that $\frac{\partial^2 x_i}{\partial b_j \partial b_\ell} = 0$ for all $1 \le i, j, \ell \le k$.

Subjective utility is given by

$$U^s(b) = u(x(b)) - p[e+b]' x(b),$$

where $x(b)$ solves

$$\nabla u(x(b)) = \begin{bmatrix} u_1(x(b)) \\ \vdots \\ u_k(x(b)) \end{bmatrix} = p[e+b].$$

Taking derivative with respect to b, and denoting the Hessian of u by H, we have

$$H\frac{\partial x}{\partial b} = H \begin{bmatrix} \frac{\partial x_1}{\partial b_1} & \frac{\partial x_1}{\partial b_2} & \cdots & \frac{\partial x_1}{\partial b_k} \\ \frac{\partial x_2}{\partial b_1} & \frac{\partial x_2}{\partial b_2} & \cdots & \frac{\partial x_2}{\partial b_k} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial x_k}{\partial b_1} & \frac{\partial x_k}{\partial b_2} & \cdots & \frac{\partial x_k}{\partial b_k} \end{bmatrix} = \begin{bmatrix} p & 0 & \cdots & 0 \\ 0 & p & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & p \end{bmatrix} \implies \frac{\partial x}{\partial b} = pH^{-1}.$$

Normative utility is given by

$$U^n(b) = u(x(b)) - pe'x(b)$$

The change in consumer surplus from changing bias by $\Delta b = \begin{bmatrix} \Delta b_1 \\ \vdots \\ \Delta b_k \end{bmatrix}$ is given by

$$\begin{aligned}
U^n(b + \Delta b) - U^n(b) &\approx \frac{\partial U^n}{\partial b}\Delta b + \frac{1}{2}(\Delta b)'\frac{\partial^2 U^n}{\partial b^2}\Delta b \\
&= \frac{\partial U^n}{\partial x}' \frac{\partial x}{\partial b}\Delta b + \frac{1}{2}(\Delta b)'\frac{\partial x}{\partial b}' H \frac{\partial x}{\partial b}\Delta b \\
&= [\nabla u(x(b)) - pe]' \frac{\partial x}{\partial b} \cdot \Delta b + \frac{1}{2}p^2(\Delta b)'H^{-1}HH^{-1}\Delta b \\
&= p^2 b' H^{-1}\Delta b + \frac{1}{2}p^2(\Delta b)'H^{-1}\Delta b.
\end{aligned}$$

It follows that if $\Delta b = -b$, the change in consumer surplus is equal to

$$-p^2 b' H^{-1} b + \frac{1}{2}p^2 b' H^{-1} b = -\frac{1}{2}p^2 b' H^{-1} b$$

which is positive if H is negative definite.

A3 Appliance-Level Electricity Use Measurement

In this section, we describe the commercial nonintrusive appliance load monitoring (NALM) algorithm, which is used in our study to determine appliance-level consumptions from high-frequency measurements of total electricity consumption of a household. Based on Hart (1992), we first introduce the general approach of NALM algorithms and then present the structure of the algorithm employed in our study (see Gupta et al. 2017 for details).

NALM algorithms exploit that appliances are typically wired in parallel, so that the power they consume is additive. The fact that appliances are switched on and off creates distinct patterns in high frequency data, which can be used to decode appliance-level consumptions. This decoding process is simplified by the fact that every appliance has a distinct signature during use, i.e. a characteristic pattern of the power it consumes. For example, washing machines use different amounts of power when they heat water, wash, and spin. NALM algorithms represent appliances as so-called finite state machines (FSMs), i.e., model appliances as having a finite set of states (e.g. off, heating, washing, spinning) and transitions between states (e.g. off $\rightarrow$ heating $\rightarrow$ washing $\rightarrow$ spinning $\rightarrow$ off). These FSM models are then mapped with observable shifts in electricity usage to determine appliance-level consumptions. While the methodology has already been proposed almost 30 years ago (e.g. Hart 1992), the mapping between FSMs and empirical transitions has been facilitated by recent advances in machine learning.

The structure of the NALM algorithm used in our study is depicted in Figure A1. A metering device records both the electric power consumed and the voltage at a high-frequency (in our case, every few seconds), thus measuring the "whole house composite load signal". This signal is analysed in order to detect so-called transitions in the data, i.e., changes in consumption levels.

A core element of NALM algorithms is a signature repository, which collects appliance signatures. To construct this repository, the algorithm uses a comprehensive collection of electrical load signature patterns of common appliances. For example, the load signature of an electric clothes dryer typically consists of three states (off, high heat, cool down) and of typical power consumptions for each of these states (e.g. 0 W, 4500-6000 W, 200-300 W, respectively). Another input is the non-electric signature repository which includes typical behavioural parameters of appliance usages (e.g. that a clothes dryer is typically used for 30-75 min). Based on these inputs, the household specific signature repository is constructed as follows. First, the NALM algorithm uses methods from cluster analysis to define clusters of shifts in electricity consumption. In a subsequent step, it classifies these clusters by comparing them to the typical states and transitions of a particular appliance. This classification step is typically performed via supervised machine learning techniques based on training data.

In a subsequent step, a load dis-aggregator uses the whole house composite load signal as well as the signal repository to decompose the entire signal into appliance-specific consumptions. In our case, load disaggregation was performed once a day, so that households could access appliance-level information always on day following appliance usage.

Appliances leave a distinct pattern in high-frequency electricity consumption data, which allows to determine the start and end date of an appliance, as well as the electricity consumed by it. For the dryer, for example, it is easy to spot the pattern of a long heating period after switching on the appliance, followed by an iteration between periods for letting cool down the laundry and heating it up again.

A3.1 Appliance-Level Feedback in the App

In this subsection, we detail how appliance-level feedback is provided to participants in our study group D. Based on the appliance-use events, we determine participants' monthly electricity consumption for every appliance category. To facilitate the assessment of appliance consumptions, we also use participants' electricity prices to calculate monthly operating cost by appliance. Furthermore, we transform monthly appliance-level consumptions into an appliance score that informs households about their usage intensity.[16] In the app, households can click on a button that provides a detailed description of the meaning of the appliance scores. The score is 100 if a household's appliance use is very low and 0 if it is very high, compared to typical usage behaviours and energy intensities of the respective appliance. More specifically, the score is calculated as follows: Appliance Score = 100 $\times$ (Monthly Appliance Consumption - $\text{Bench}_{\text{low}}$)/($\text{Bench}_{\text{high}}$ - $\text{Bench}_{\text{low}}$), where $\text{Bench}_{\text{low}}$ and $\text{Bench}_{\text{high}}$ correspond to pre-determined benchmark values for high and low appliance uses, respectively. We construct these benchmarks from survey data on typical appliance uses as well as product data sheets on the technical efficiency of appliances currently used in German households (for details, see Table A4 in the Appendix).

In Figure A5, we display the distributions of the appliance scores by appliance category. As we determined the appliance score benchmarks prior to the experiment, assessing the range of the appliance scores serves to evaluate the plausibility of the detected appliance use events. For all appliance categories, the vast majority of appliance scores lies between 0 and 100, which supports the credibility of the disaggregation. For the categories *Dishwasher*, *Dryer* and *Oven*, there is bunching at indices of 100, which indicates that some participants have not used these appliances at all in some months.

A3.2 Plausibility checks

To test the plausibility of the appliance-level measure, we benchmark the appliance-level measurements by comparing our data from 2017 with the average appliance uses in Germany, which is only available for 1996 and 2011. As Figure A6 in the Appendix shows, the percentage of electricity used for cooking (9.6%) and for washing, drying, and dish-washing (9.7%) aligns with German averages in 1996 and 2011 (about 9.6 – 9.8% and 10.4 – 12.4%, respectively). In our study, refrigeration accounts for 9.9%, which is less than German averages for 1996 and 2011 (22.6% in 1996 and 16.7% in 2011). This divergence likely reflects a gradual increase in energy efficiency of refrigerators and freezers over time, not least owing to ever increasing minimum standards (see, e.g., Andor et al. 2020b). The percentage of the category *Other Appliances* amounts to 71%, which is slightly larger than the German averages (57.5% in 1996 and 61.1% in 2011). This deviation is likely driven by the general trend that households use more

[16] After consulting with the app designers, we denoted this score as an "efficiency score", as this term has intuitive appeal to an average household.

electric devices, such as smart TVs, computers, smartphones, and robotic vacuum cleaners. By contrast, air conditioning is not prevalent in German households and thus cannot explain that increase.

A4 Matched Control Group

To identify the overall conservation effect, we additionally obtain data from a non-experimental sample of smart meter households that are served by the same utility. These households have agreed to report their electricity consumption to the grid operator who uses the data to forecast load profiles. We obtain smart meter data in 15 minute intervals for 577 households, starting from November 1, 2016, which is when our field test started. This data allows us to identify the effect sizes for all experimental conditions relative to obtaining no feedback at all. To ensure that observable household characteristics are balanced across our experimental and non-experimental sample, we select a control group using a propensity matching method.

As the left panel of Figure A7 in the Appendix shows, the baseline consumption in the non-experimental sample is slightly larger than in the experimental sample. To account for such differences, we follow a matching approach to determine the subset of control households that we use in our analyses, denoted henceforth as matched control (MC) households. For every participant in the Aggregate Feedback group, we determine the nearest neighbour in the non-experimental sample by implementing a 1:1 matching algorithm without replacement, based on two covariates. First, we match on the average per day electricity consumption in the baseline period. Second, we control for differences in the timing of billing periods by matching on the end day of the billing period before the start of the field test. As a result, we obtain a group of 140 matched households.

The right panel of Figure A7 shows that, after matching, participants in the A and in the MC group have about the same baseline electricity use distribution. In Table A8, we additionally assess the balance in terms of further billing information. We find that the average end date of the bill is not statistically different for both groups, as expected from matching. Yet, the average start of the billing period starts about one month earlier in the Aggregate Feedback (A) group, compared to the matched non-experimental observations. To account for such differences in the baseline billing period, we include month-of-baseline fixed effects in our post-matching regressions.

A5 Secondary Analyses and Robustness Checks

In the following, we describe in more detail our secondary analyses and robustness checks.

A5.1 Heterogeneity in Treatment Effects by Baseline Consumption

To investigate treatment effect heterogeneity by baseline consumption in more detail, we estimate two equations. First, we estimate the following equation by OLS, using only our experimental sample:

$$Y_{it}^{norm} = \alpha Y_i^{b,dm} + \beta A_i Y_i^{b,dm} + \nu_t + \mu_w^b + \epsilon_{it},$$

where $Y_i^{b,dm}$ denotes the baseline consumption of household i, expressed as a percentage of the average daily consumption in the A group. We also demean this variable, so that we can interpret $\hat{\beta}$ as the average treatment effect at the mean of baseline consumption. In addition, we estimate the following equation using the experimental and the matched control groups:

$$Y_{it}^{norm} = \alpha Y_i^{b,dm} + \beta_0 EC_i + \beta_1 EC_i Y_i^{b,dm} + \gamma_0 D_i + \gamma_1 D_i Y_i^{b,dm} + \nu_t + \mu_w^b + \epsilon_{it}.$$

Using the experimental sample, the estimates for the interaction term between the disaggregation dummy D and baseline electricity consumption amounts to -0.66, but is not statistically significant (Panel a of Table A9). When using data from the experimental sample and matched control observations, we find that the interaction effect reaches -0.124 and is statistically significant at the 1% level (Panel b of Table A9). Furthermore, we cannot reject the null hypothesis that both the main effect of EC and its interaction with the baseline electricity use are zero (F-test stat.: 1.35, p-value: 0.26), while we can reject the corresponding null hypothesis for the disaggregation groups D (F-test stat: 7.47, p-value: 0.001).

A5.2 Treatment Effects by Hour-of-the-Day

In this subsection, we exploit the high granularity of our data to investigate how treatment effects vary by hour-of-the-day. This analysis allows us to explore the timing of households' responses to appliance-level feedback. Using households from our experimental sample, we estimate the following equation:

$$Y_{ith}^{norm} = \alpha Y_i^b + \sum_{k=1}^{24} \beta^k 1(k = h) D_i + \nu_t + \mu_h + \mu_m^b + \epsilon_{ith}, \tag{46}$$

where $1(\cdot)$ is the indicator function and all variables are defined as in Equation (1), except that we now investigate the hourly electricity consumption of participant i on day t and hour h, and additionally include 24 fixed effects μ_h for every hour of the day. We normalize our outcome variable by the average hourly consumption in the Aggregated Feedback (A) group, so that our estimates $\hat{\beta}^h$ capture the ATE in hour h, expressed as a percentage of the average consumption in that hour. Again, we cluster standard errors at the household level.

Figure A11 shows that treatment effects are large during late morning hours and late evening hours. During these hours, they reach about 10% of the average control group consumption. Furthermore, we find that the magnitude of electricity conservation cannot be predicted by baseline electricity consumption levels alone. Electricity reductions are particularly strong in the late morning hours between 8 a.m. and 1 p.m., which coincide with large electricity consumption levels. However, during 4 and 8 p.m., consumption levels are similar, but households save considerably less. We also detect strong savings in late evening hours between 9 and 11 p.m., when consumption levels are rather low.

Using households from the A and MC groups, we also explore the timing of the electricity savings from aggregate feedback (by estimating (46) and replacing the dummy variable D_i with the dummy variable A_i that equals one for households in the A group). As shown in Figure A10, we find that the electricity savings occur during the evening hours, between 6 and 8 p.m.

A5.3 Appliance-Level Treatment Effects by Hour of the Day

To identify the hourly average treatment effects at the appliance-level, we estimate the following model separately for every appliance:

$$Y_{ithj}^{norm} = \alpha Y_i^b + \sum_{h=1}^{24} \beta_j^h D_i + \nu_t + \mu_h + \epsilon_{ihdj}.$$

In Figure A12, we show how appliance-level consumptions change in response to appliance-level feedback over the hours of a day. For dish-washers, dryers, and washing machines, we find a distinct pattern that savings occur only during the day, between 7 a.m. and 15 p.m., which coincides with typical usage patterns of these appliances. By contrast, consumption reductions in the category *Other Appliances* occur particularly during late morning hours, as well as during late evening hours, between 8 p.m. and 4 a.m. As the categories *Refrigeration* and *Always-On* are measured daily, we cannot estimate hourly treatment effects for them.

A5.4 Appliance-Level Beliefs

We conducted surveys before and after the start of our intervention to elicit participants' socio-demographic characteristics, household characteristics, and beliefs. The first survey took place in November 2016 prior to the start of the field test, followed by two additional surveys in March 2017 and July 2017.

The survey that we conducted prior to the field test allows us to assess the extent to which households misjudge their aggregate and appliance-level electricity use. We elicited households' annual electricity consumption beliefs and compare them to their annual consumption values from the latest annual bill prior to the field test. This comparison does not provide any evidence that households underestimate their aggregate electricity consumption. On average, annual consumption beliefs are virtually on par with baseline consumption levels (Panel b of Table A1). In addition, we find that beliefs closely match the actual consumption values. A one-unit increase in baseline electricity use is associated with an increase in baseline beliefs by 0.93 units (Std. Err.: 0.026; see Table A7 and Figure A13 in the Appendix). As shown in Byrne et al. (2020), this estimate is a measure of the consumers' misperception of aggregate consumption. Hence, our finding that it is only slightly below one indicates that misperceiving aggregate consumption is unlikely to cause excessive resource use in our context.

To assess appliance-level beliefs, our baseline survey contained a wattage ranking task that asked households to assess the wattage of a laptop, dish-washer, tumble dryer, and mobile heater relative to a 100 watt lightbulb on a five point Likert scale ("much lower", "lower", "about the same", "higher", "much higher").[17] We find that more than half of all respondents make at least one mistake in assessing appliance-level wattage. Furthermore, respondents make mistakes in about 11% of all the binary appliance-level wattage comparisons that are implicitly incorporated in the task. Both findings show that households' knowledge of appliance-level energy intensities is limited, as formalized by our model.

To investigate the mechanisms that underlie households' responses to appliance-level feedback, we first explore whether treatment effects are particularly large for households who make mistakes in the wattage ranking task. This test is closely linked to our model, which predicts that households who misjudge the relative wattage of appliances should respond most strongly to the provision of appliance-level information. We define two groups of households: those who made at least one mistake in the wattage ranking task and those who did not. We then estimate our main specification (Equation 4) and interact the dummy variable D with both group indicators, which yields the average treatment effect of appliance-level information in these groups. As shown in Column (1) of Table A16a, we find that the effect size amounts to -4.3% for households who make at least one mistake and is statistically significant at the 5% level. For households with no errors in the wattage ranking task, the effect size amounts to only -1.4% and is not statistically significant at any conventional level. In Column (2), we also consider a third group of households who did not indicate appliance-level beliefs in the baseline survey. The effect size that we estimate for this group of households is even larger than for those households who indicate beliefs and make at least one mistake. As not expressing be-

[17] Inspired by Attari et al. (2010), the wattage ranking task also included three further appliances: desktop computer, hifi system, and air conditioning. Depending on the configuration of these devices, their wattage varies to an extent that prevents us from establishing a clear wattage ranking. We thus exclude these devices from our analysis.

liefs may indicate particularly poor knowledge of appliance wattages, this finding is consistent with our previous results.

In addition, we explore the intensive margin of making mistakes by interacting the dummy variable D with the share of mistakes in all binary comparisons implied by the wattage ranking task (Column 3). The estimate of the main effect (D) shows that households who make no mistakes save only about 1.2% in response to appliance-level feedback. By contrast, we find that an increase in the share of mistakes from 0 to 1 increases conservation effects by 15.6 percentage points (p-value: 0.069), which supports the idea that households with poor knowledge of appliance wattage have particularly large treatment effects. Furthermore, the regression results from Column 3 demonstrate that households who make mistakes in the wattage ranking task tend to have higher consumption levels overall. Taken together, these findings suggest that lack of energy-related knowledge among households with high consumption levels could explain why many previous feedback studies have detected particularly large treatment effects for these households (e.g., Allcott, 2011; Andor et al., 2020a; Tiefenbeck et al., 2018).[18]

As a complementary test for the role of beliefs, we explore whether participants' appliance-level beliefs align more closely with the appliance-level uses that we measure during the experiment. We elicited appliance-level beliefs in a baseline survey that took place prior to the study period (November 2016) and an end-line survey towards the end of the study period (July 2017). In particular, we asked participants to estimate their monthly electricity consumption for the appliance categories always-on, washing machine, dryer, refrigeration, and dishwasher. A limitation of our data is that we observe appliance-level uses only once our field experiment started. Hence, we cannot assess the accuracy of households' appliance-level beliefs before obtaining appliance-level feedback. Yet, we can test whether beliefs during the field tests align more closely with measured appliance-level uses for households in the treatment groups T_1-T_4 than for households in the Aggregate Feedback (A) group.

To circumvent difficulties that may arise from noisy appliance-level belief measurements, we first translate all consumption beliefs into ranks.[19] We assign rank 1 if a participant believes that the monthly consumption of an appliance was highest among all of his or her appliances. Similarly, rank 2 corresponds to a belief that the appliance consumption occupies the second rank, etc. We also calculate the same ranks based on the appliance-level data that we can observe. This allows us to calculate a rank difference as the absolute difference between the ranks implied by participants' belief and those based on the appliance-level data during the study period. Averaging over these absolute differences across all appliances yields a measure of the accuracy of beliefs that is robust to differences in the unit of measurement used by participants when expressing their beliefs. For this measure to be comparable, we drop all

[18]In our study, we also find evidence for such treatment effect heterogeneity (see Appendix Section A5.1 for details).

[19]The elicitation of appliance-level energy consumption beliefs is subject to vivid controversy and a methodological consensus has not been reached so far (Frederick et al., 2011; Attari et al., 2010, 2011). To help consumers who are not familiar with energy consumption units, some researchers provide reference points and inform households about the energy consumption of a reference appliance prior to asking participants about energy consumption beliefs (Attari et al., 2010). While this approach can reduce excessive variance in participants' answers, providing a reference point has been shown to also bias belief estimates towards that reference point. In addition, changing the unit of measurement from watts to kilowatts, for example, can induce framing effects that also bias belief elicitation (Frederick et al., 2011). For these reasons, we chose not to provide reference points.

participants who have not stated beliefs for all appliances in one of the two surveys. The mean rank difference is zero if a consumer is correct about the rank of all appliances and can reach up to 3 if the estimated consumption ranks are exactly opposite to the ranks from our appliance-level measurements.

Table A16b displays the average absolute rank difference for the baseline survey that we conducted prior to the experiment (B) and the survey after the core study period (E), as well as the change in the rank difference (E-B). The baseline rank difference amounts to about 1.8 for all experimental groups. It reflects poor knowledge of energy intensities as it is only slightly lower than the expected rank difference of randomly determined ranks, which amounts to $70/36 \approx 1.94$. For the *EC* group, we cannot reject the null hypothesis that the mean differences in the rank difference (E-B) equals zero at any conventional significance level. This finding is consistent with the absence of an conservation effect for that group (for details, see Section A4 in the Appendix). By contrast, for each of the treatment groups T_1-T_4 with appliance-level feedback, we observe a decrease in the absolute rank difference that is statistically significant at the 5% level. Accordingly, participants in these groups adjusted their beliefs in response to obtaining appliance-level feedback.

A5.5 Sub-Treatment Arms

In this subsection, we explain the four sub-treatment arms that all receive disaggregate information (and thus form the Disaggregate Feedback group in the main text).

In addition to the app functionalities that participants in the Aggregate Feedback group can use (see Section 3.1 for details), participants in our first sub-treatment arm T_1 have access to an additional app page that provides feedback on appliance-level usages, cost and appliance scores. In the three sub-treatment arms T_2-T_4, participants receive the same appliance-level feedback and are additionally invited to take part in appliance challenges. With these challenges, we test whether complementary interventions increase the effectiveness of appliance-level feedback. The challenges start at the beginning of the second month after installation of the app and require participants to increase one of the appliance scores by as much as they can within a month. At the end of the month, the change in the appliance score relative to the previous month is evaluated as follows.

In sub-treatment arm T_2, participants obtain 1 EUR per appliance score improvement, capped at a maximum of 20 EUR per monthly challenge. This treatment is motivated by studies that have found stronger conservation effects when monetary incentives are provided (e.g., Ito et al. 2018). In T_3, participants receive a ranking that compares their score improvement with those of other study participants, but do not obtain a financial reward. A participant within the top percentile of monthly score improvements is classified as rank one, and similarly for all other percentiles. With this treatment, we test the impact of relative performance feedback, which has been found to be effective in educational (Azmat and Iriberri, 2010; Tran and Zeckhauser, 2012) and workplace settings (Mas and Moretti, 2009; Blanes i Vidal and Nossol, 2011). In T_4, we implement the same ranking, but reward participants according to their rank: rank one translates into 10 EUR, rank two into 9.9 EUR, etc., and rank 100 into 0 EUR.

This treatment allows us to explore the effectiveness of combining relative performance feedback with monetary incentives.

Participants take part in a maximum of five challenges. The first challenge is always targeted towards the appliance with the lowest appliance score, followed by the appliance with the second-lowest score, etc. If less than five appliances are detected for a household, challenges can target the same appliances more than once, given that all other appliances have been targeted already. At the end of our study, participants in T_2 and T_4 receive an Amazon voucher to the amount of their earnings from taking part in the challenges. We designed our reward scheme to yield similar average payments in both treatment groups. The average realized payments per monthly efficiency challenge amount to 6.3 EUR in T_2 and to 4.5 EUR in T_4.[20] The timing in our experiment is as follows: In the first month after app installation, participants in T_1–T_4 obtain appliance-level feedback but challenges have not yet started. In the months 2-6, participants in T_2-T_4 take part in challenges. After month 6, challenges do not occur any more, but T_1–T_4 participants continue to receive appliance-level feedback.

We disentangle the differential impact of the app elements by estimating the following equation:

$$Y_{it}^{norm} = \alpha Y_i^b + \beta_1 D_i + \beta_2 M_i + \beta_3 R_i + \beta_4 M_i R_i + \nu_t + \mu_w^b + \epsilon_{it}, \tag{47}$$

where the scalar D_i equals one if a household is in any one of the four treatment groups $(T_{1i}, T_{2i}, T_{3i}, T_{4i})$, so that $\hat{\beta}_1$ identifies the conservation effect of providing appliance-level feedback, compared to providing aggregate feedback only. Furthermore, M_i equals one for the treatment groups T_{2i} and T_{4i}, where participants receive monetary rewards for saving electricity. Similarly, R_i equals one for the treatment groups T_{3i} and T_{4i}, where participants obtain information on their savings relative to those of other participants. All three groups T_2-T_4 also receive appliance-level feedback. Hence, the parameter estimates $\hat{\beta}_2$ and $\hat{\beta}_3$ identify by how much the effectiveness of appliance-level feedback changes when monetary incentives and rank information are provided additionally. We also interact M_i and R_i to test whether the effectiveness of monetary incentives increases when they are tied to a relative ranking rather than an absolute appliance score improvement. This interaction effect is identified by parameter β_4.

Our results from Panel b) of Table A11 show that the provision of appliance-level feedback in T_1-T_4 is the main driver of the electricity conservation we observe. Households that obtain such feedback reduce their consumption by almost 5 percent compared to households with aggregate feedback (Column 3). In Column (4), we test whether the provision of monetary incentives and rankings intensify the response to appliance-level feedback. We find that the point estimates are close to zero and not statistically significant at any conventional level. Hence, neither monetary incentives nor rankings trigger higher electricity savings compared to appliance-feedback alone. Furthermore, we do not find support for the conjecture that monetary incentives become more effective when information about participants' rank is also provided, as shown by the small and statistically insignificant interaction effect between M and R in Column (5). Taken together, our evidence suggests that appliance-level information alone

[20]The average payment in group T_4 does not exactly equal 5 EUR as we have less than 100 challenge participants in that group for some months.

leverages its full potential and that there are no significant complementarities with monetary or social incentives. As we do detect no differences between the ATEs of the treatment groups $T_1 - T_4$, we pool them and use the disaggregation dummy D for our analyses in the main text.

A5.6 Conservation Effects after Core Study Period

After our core study period of 6 months, participants continued to have access to the app, but with limited functionality in treatment group $T_2 - T_4$. In particular, participants were not invited to take part in efficiency challenges any longer, but still received appliance-level consumption feedback.

To investigate the treatment effects after the core study period, we estimate Equation (4) and (47), but restrict the sample to the time period from month 7 of the field test onwards.

Panel a) of Table A12 gives the average conservation effect relative to the EC group and shows that the point estimate for the treatment effect in the disaggregation groups $T_1 - T_4$ amounts to -1.4%, but is not statistically significant at any conventional level. When we restrict our analysis to a balanced panel, the treatment effect amounts to -3.3%, but is not statistically significant at any conventional level. Using the non-experimental sample to identify the total conservation effect (Panel b of Table A12), we find that disaggregate feedback yields a persistent reduction by -4.7% when using the full sample (Column 3) and of -2.7% when restricting our sample to a balanced panel of households (Column 4).

We test for differences in the average treatment effect between the core study period and the period thereafter by estimating the following regressions based on our experimental sample, as well as on our experimental sample and our matched control group, respectively:

$$Y_{it}^{norm} = \alpha Y_i^b + \tau\ AfterCSP_t + \beta_1 D_i + \beta_2 D_i \cdot AfterCSP_t + \nu_t + \mu_m^b + \epsilon_{it},$$

$$Y_{it}^{norm} = \alpha Y_i^b + \tau\ AfterCSP_t + \gamma_1 EC_i + \gamma_2 EC_i\ AfterCSP_t + \delta_1 D_i + \delta_2 D_i\ AfterCSP_t + \nu_t + \mu_m^b + \epsilon_{it},$$

where $AfterCSP_t$ equals one if day t occurs after the beginning of study month 7 and zero otherwise. To avoid that differences in average treatment effects arise from changes in the sample composition over time, we use a balanced panel for both regressions. The estimate $\hat{\beta}_2$ identifies the difference in the average treatment effect from appliance-level feedback relative to the Aggregate Feedback (A) group: ATE_{AfterCSP}-ATE_{CSP}, where ATE_{CSP} denotes the ATE during the core study period. The estimate δ_2 has the same interpretation, but gives the change in the ATEs relative to the matched control group, i.e., relative to obtaining no feedback at all. As shown in Table A13, we find that that $\hat{\beta}_2$ and $\hat{\delta}_2$ are positive, but not statistically significant from zero at any conventional level. While inconclusive, it is possible that the the strong seasonal decline by nearly 40% evident in Figure 2 in the main text plays a role. The decline may be driven by uses that are more responsive to disaggregate feedback. It is also possible that attrition contributed to this trend. Point estimates of the impact of disaggregation are slightly larger when we restrict the sample to observations who never experienced any data loss. This is true for the main study period (see Table A10), as well as for the treatment effect of disaggregation after the core study period (see columns (2) and (4) in Table A12).

A6 Elasticity Estimates and Sensitivity Checks

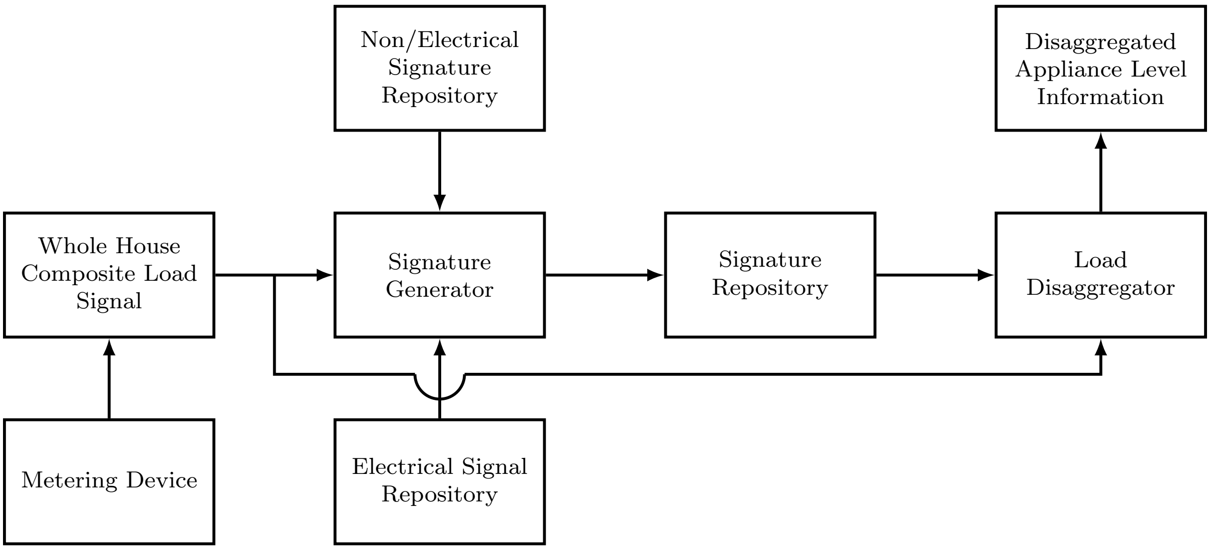
To identify η_j from our data, we estimate the following regression separately for every appliance category j:

$$ln\, y_{ij} = \eta_j \, ln\, p_i + \beta_j' \mathbf{X}_i + \epsilon_{ij},$$

where $\mathbf{X}$ denotes a vector of socio-demographic variables for household i such as its income and size, as well as the annual consumption level in the year prior to the experiment. Cross-sectional identification faces some challenges that we discuss in the following. A typical concern for identification is non-linear pricing, where marginal prices change with the level of electricity consumption. As households in our sample face constant marginal prices, non-linear pricing does not threaten the consistency of our estimates. Another concern is omitted variable bias. Electricity suppliers offer tariffs with lower marginal prices to households with higher consumption levels, which could negatively bias our elasticity estimates. To circumvent such bias, we control for baseline electricity consumption, as well as for household income and size. While controlling for additional covariates reduces concerns from omitted variable bias, it may still be present to some degree. Our estimation results can be found in Table A17. As an indirect test of the magnitude of our elasticity estimates (see Section 5), we calculate the household-level elasticity implied by our appliance-level estimates and compare it to the findings by Frondel et al. (2019). If our appliance-level elasticity estimates systematically suffered from omitted variable bias, we would expect to find that the implied household-level elasticity is biased as well. Further sensitivity checks are presented in Table A15, as also discussed in the main text in Section 5.

Supplementary Figures

Figure A1: Schematic Representation of the NALM Algorithm

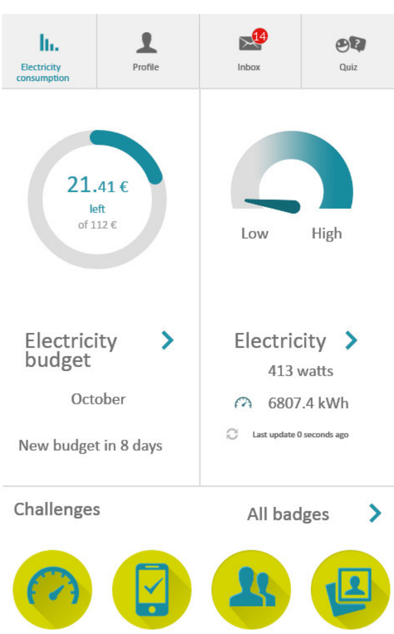


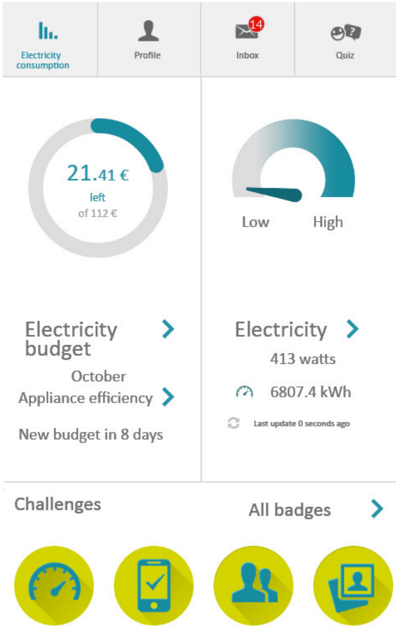
Notes: The representation is based on Gupta et al. (2017).

Figure A2: Visualization of Screens I

Start Screen (by Experimental Condition)

Aggregate Feedback (A) Disagg. Feedback (T_1)





Disagg. Feedback + Efficiency Challenges (T_2-T_4)

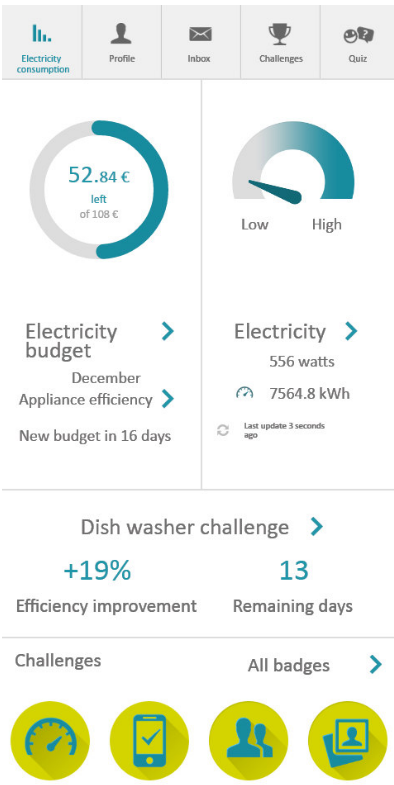


Figure A3: Visualization of Screens II

Treatment-Specific Screens

Appliance-Level Feedback (T_1-T_4)	Efficiency Challenge (T_2-T_4)

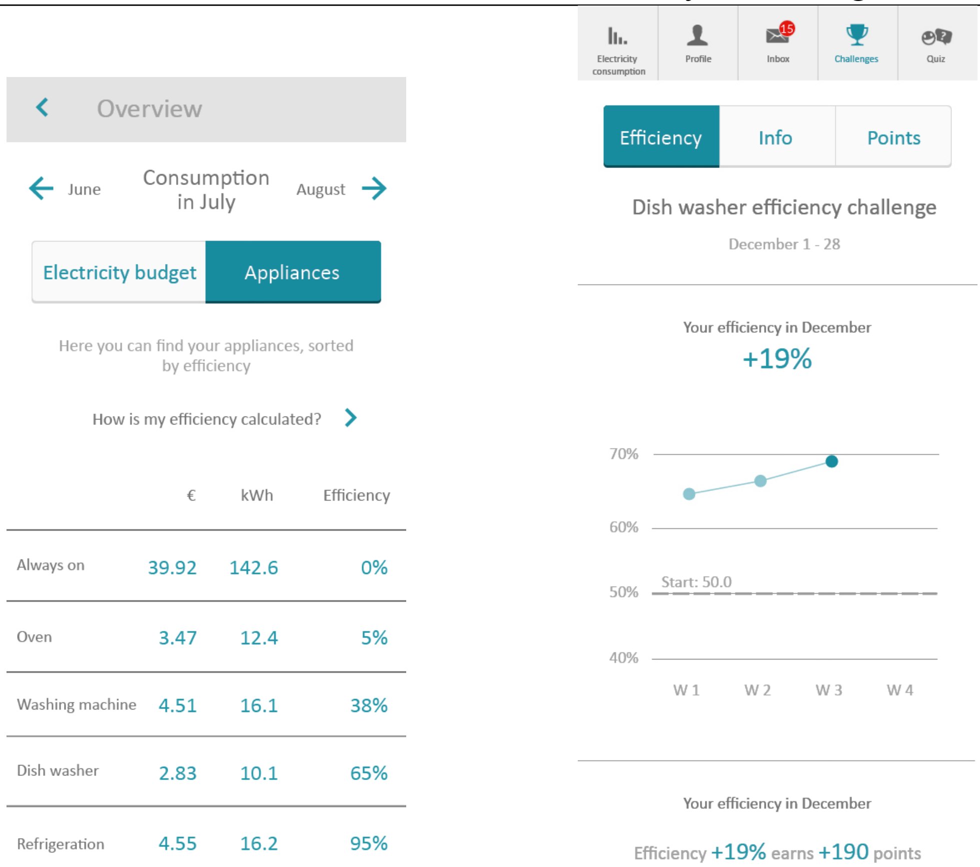
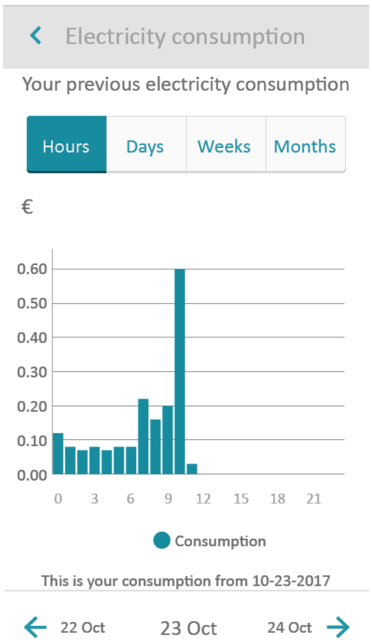
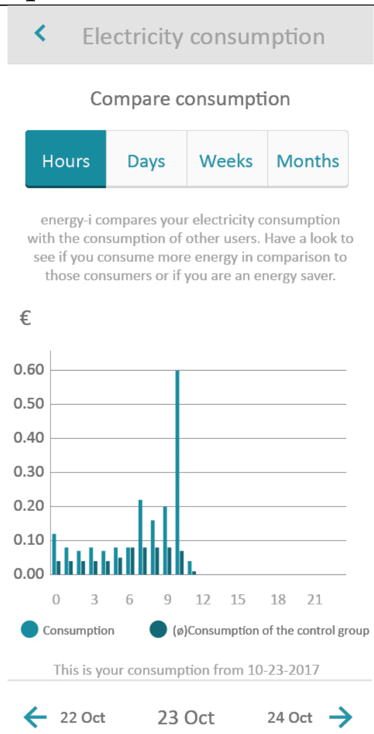


Figure A4: Visualization of Screens III

Add-On Functionalities (All Groups)

Comparison with Others	Comparison with own History





Current Power

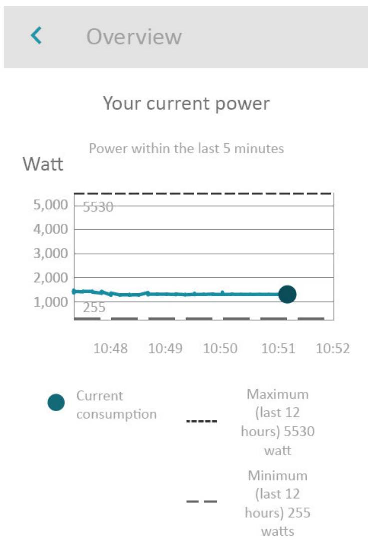
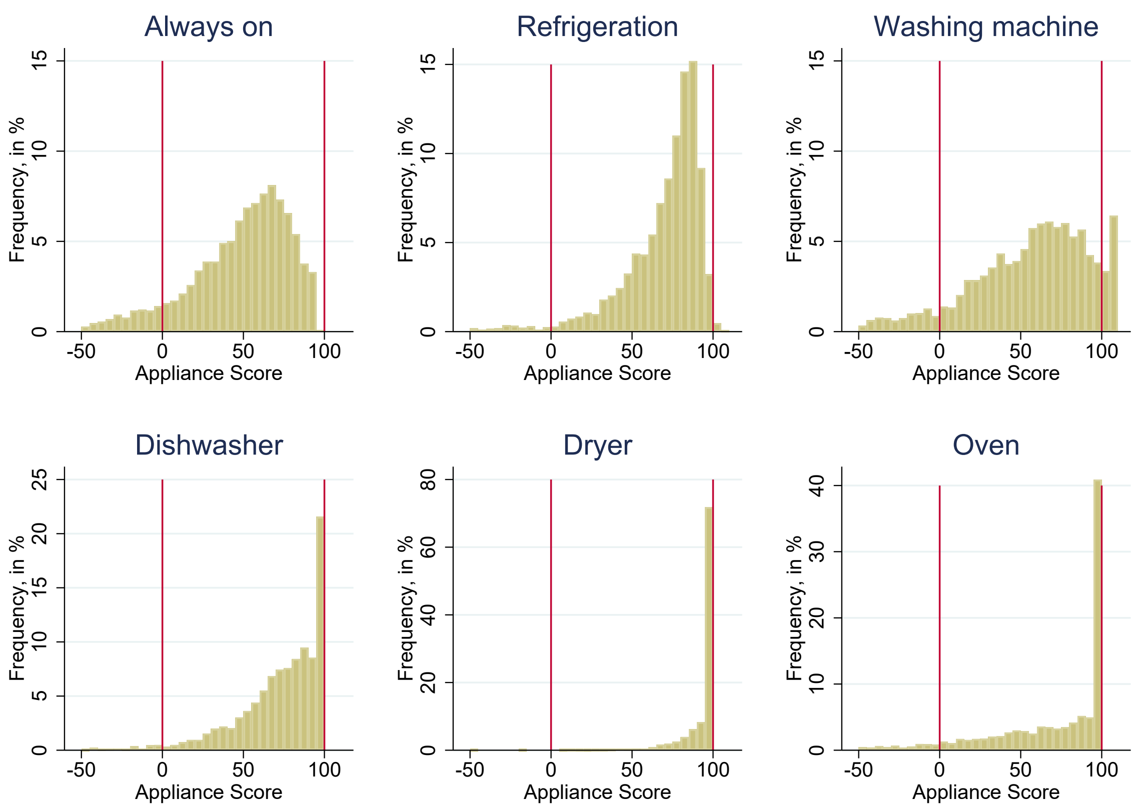
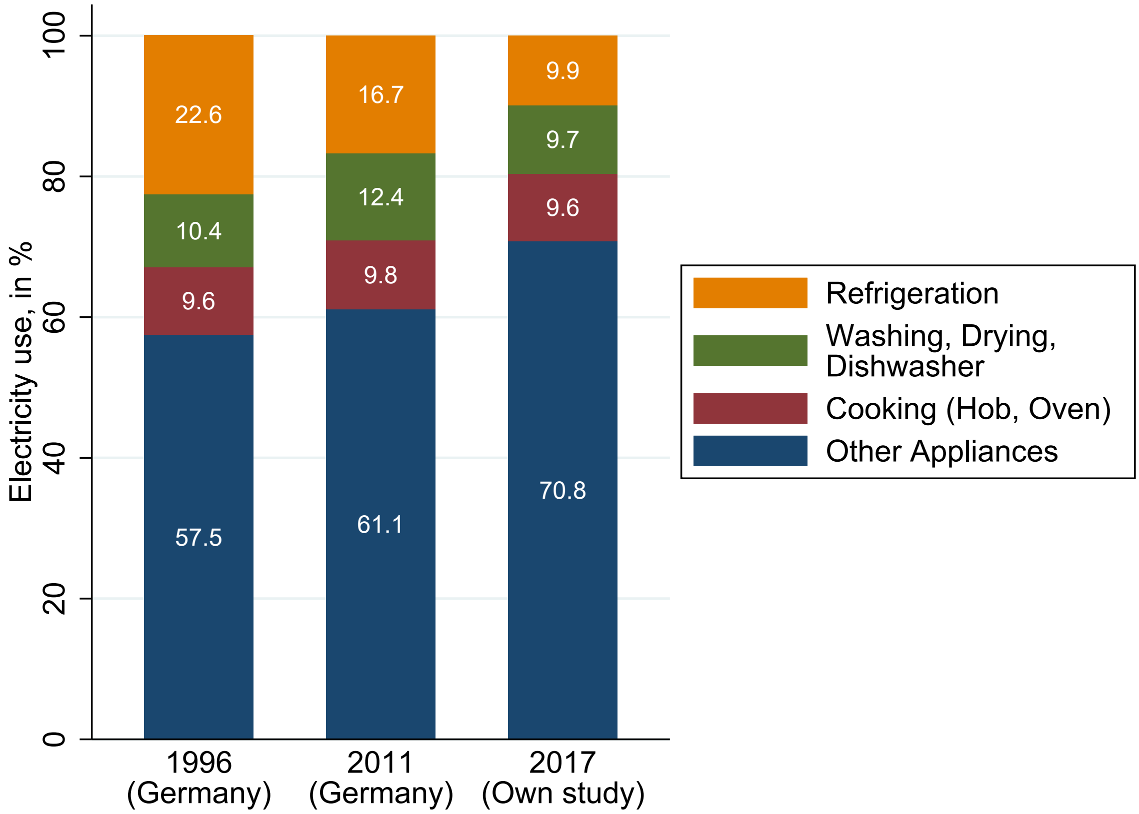


Figure A5: Distribution of Appliance Scores



Notes: Appliance Scores are calculated as follows: Appliance Score = 100 × (Monthly Appliance Consumption - $\text{Bench}_{\text{low}}$)/($\text{Bench}_{\text{up}}$ - $\text{Bench}_{\text{low}}$), where $\text{Bench}_{\text{low}}$ and $\text{Bench}_{\text{high}}$ correspond to pre-determined benchmark values for high and low appliance uses, respectively. These benchmarks are based on survey data on typical appliance uses and product data sheets on the technical efficiency of appliances currently used in German households (for details, see Table A4).

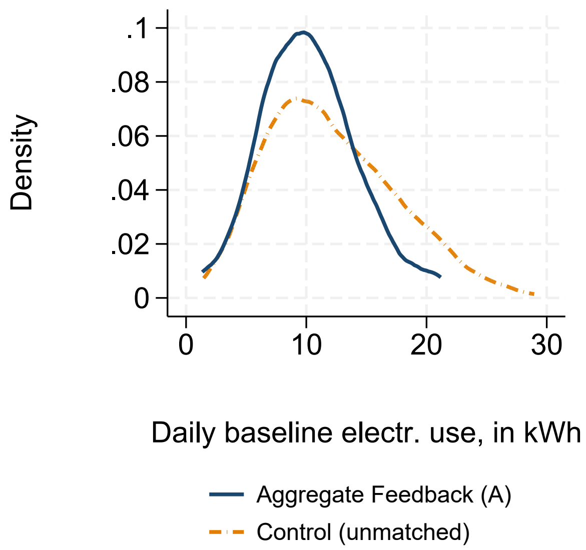
Figure A6: Decomposition of Electricity Uses, by Appliance Category



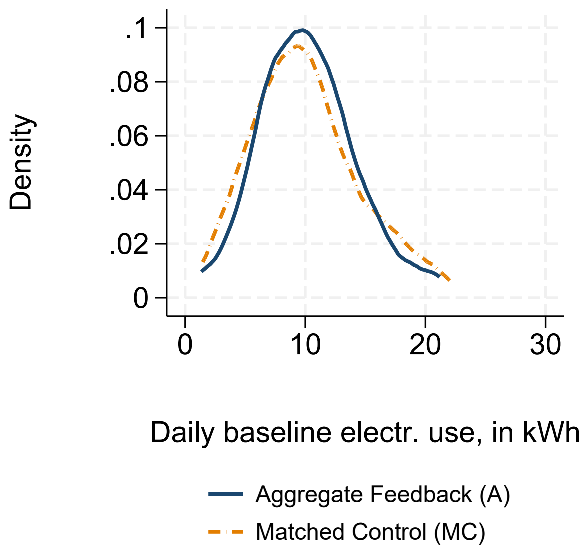
Notes: Values are expressed as percentages of total electricity consumption in a year. Values for 1996 and 2011 are drawn from BDEW (2016), Energie-Info - Stromverbrauch im Haushalt, Bundesverband der Energie- und Wasserwirtschaft. Values from our study are calculated for the Aggregate Feedback (A) group. As we do not have data on hobs, we extrapolate their consumption based on the rule-of-thumb that a hob accounts for 77.5% (75-80%) of total electricity consumption for cooking, as stated by the energy efficiency advocacy HEA (`www.hea.de/fachwissen/herde-backoefen/betriebswerte-und-energieverbrauchskennzeichnung`, last access: February 27, 2020).

Figure A7: Balancing in Terms of Baseline Electricity Consumption between Study Participants and the Non-Experimental Sample

(a) Prior to Matching



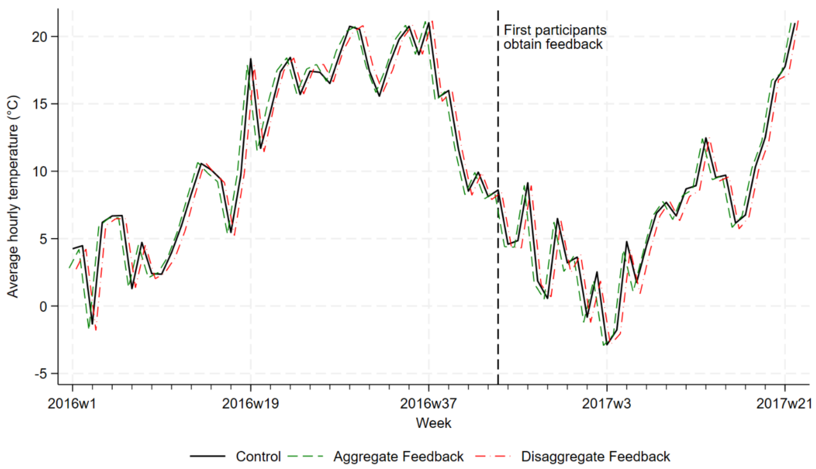
(b) After Matching



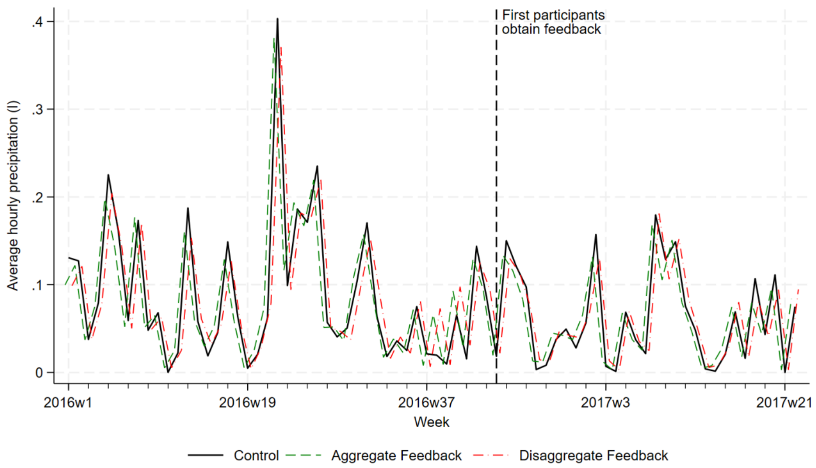
Notes: "Control (unmatched)" denotes all control group households, while "MC: Matched Control" denotes the group of households that have been matched to households in the Aggregated Feedback (A) group.

Figure A8: Seasonal Factors Across Treatment Groups

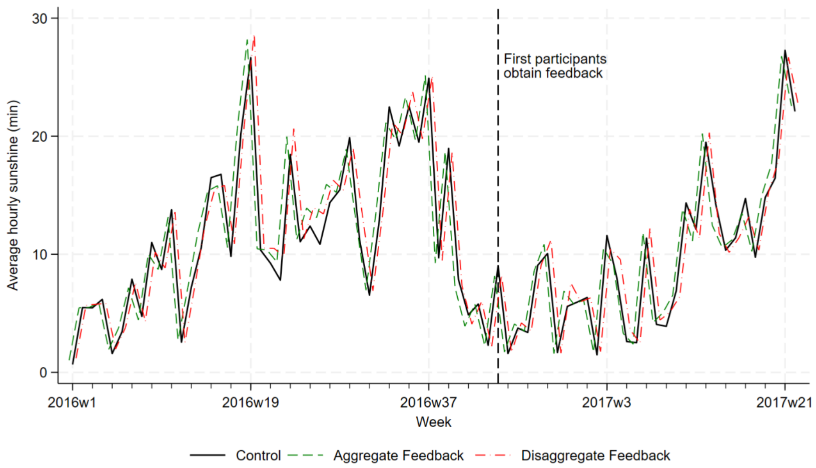
(a) Temperature



(b) Precipitation



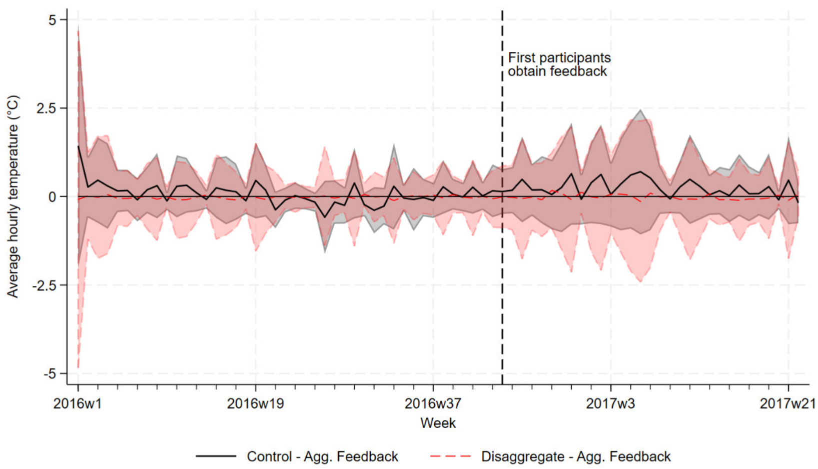
(c) Sunshine



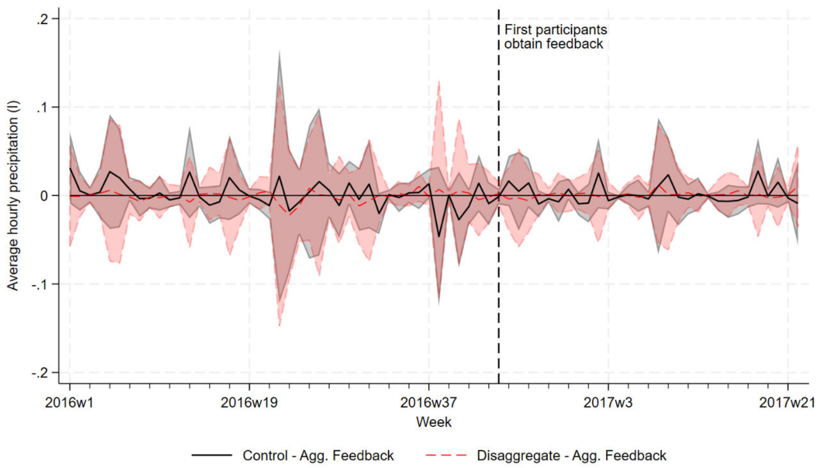
Notes: ”Control” denotes all households in the matched control group, while Aggregate Feedback and Disaggregate Feedback refer to the experimental groups.

Figure A9: Seasonal Differences between Treatment Groups

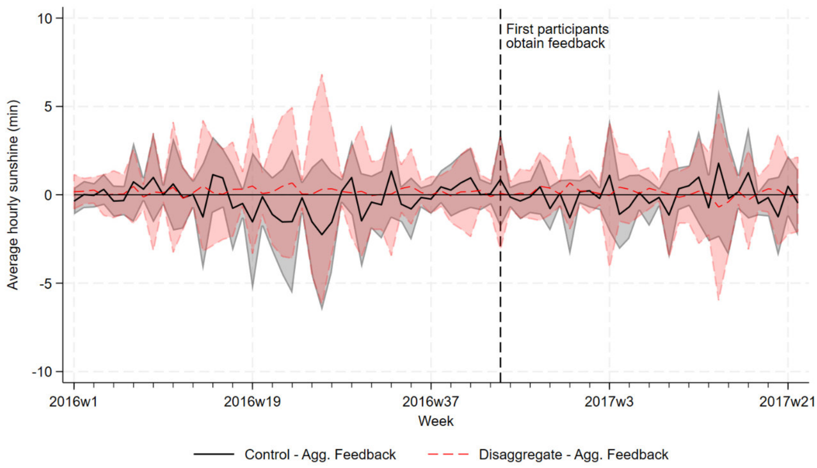
(a) Temperature



(b) Precipitation

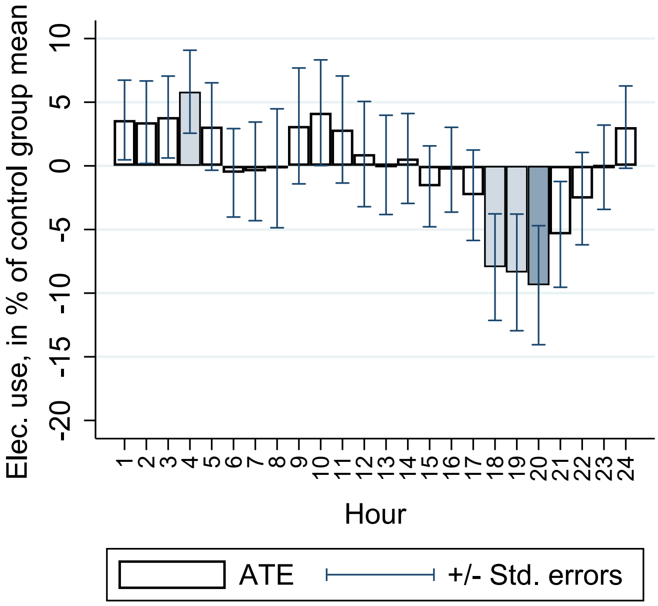


(c) Sunshine



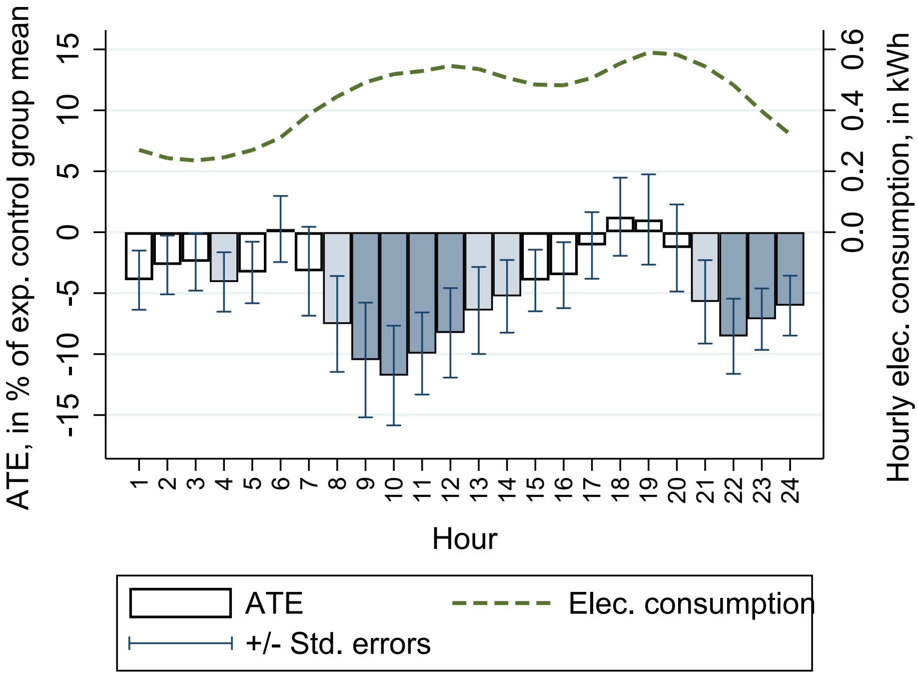
Notes: "Control - Agg. Feedback" denotes the difference between the outcome in the matched control and aggregate feedback group (likewise for Disaggregate - Agg. Feedback. Shaded areas denote 95% confidence intervals (non-parametric heteroscedasticity-consistent estimator, correcting for spatial correlation (Hsiang, 2016)

Figure A10: ATE of Aggregate Feedback (A) group on Hourly Electricity Consumption (relative to Matched Control)



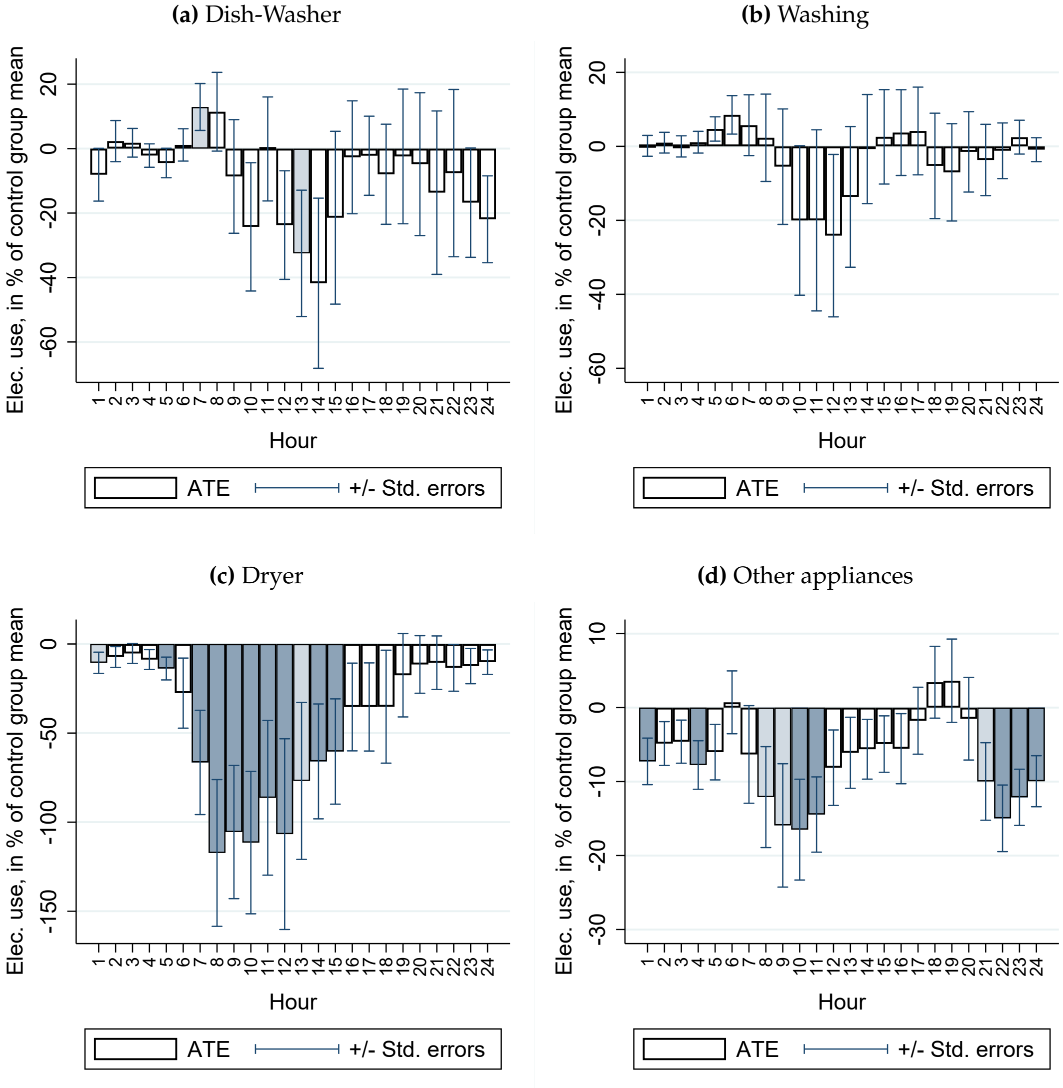
Notes: Shaded bars indicate that treatment effects are statistically significant at the 1% (blue shaded) or 5% (light blue shaded) level. Whiskers indicate a range of +/- 1 standard error (clustered at the household-match level). The outcome variable is daily electricity consumption, divided by the mean in the A group. Using participants in the MC and the A groups, we estimate the following equation: $Y_{ith}^{norm} = \alpha Y_i^b + \sum_{h=1}^{24} \beta^h EC_i + \nu_t + \mu_h + \epsilon_{ihd}$. We cannot reject the null hypothesis that all hourly point estimates are zero: $F(24, 277) = 1.15$, p-value: 0.2938.

Figure A11: ATE of Disaggregation (D) relative to Aggregate Feedback (A) group, by Hour of the Day



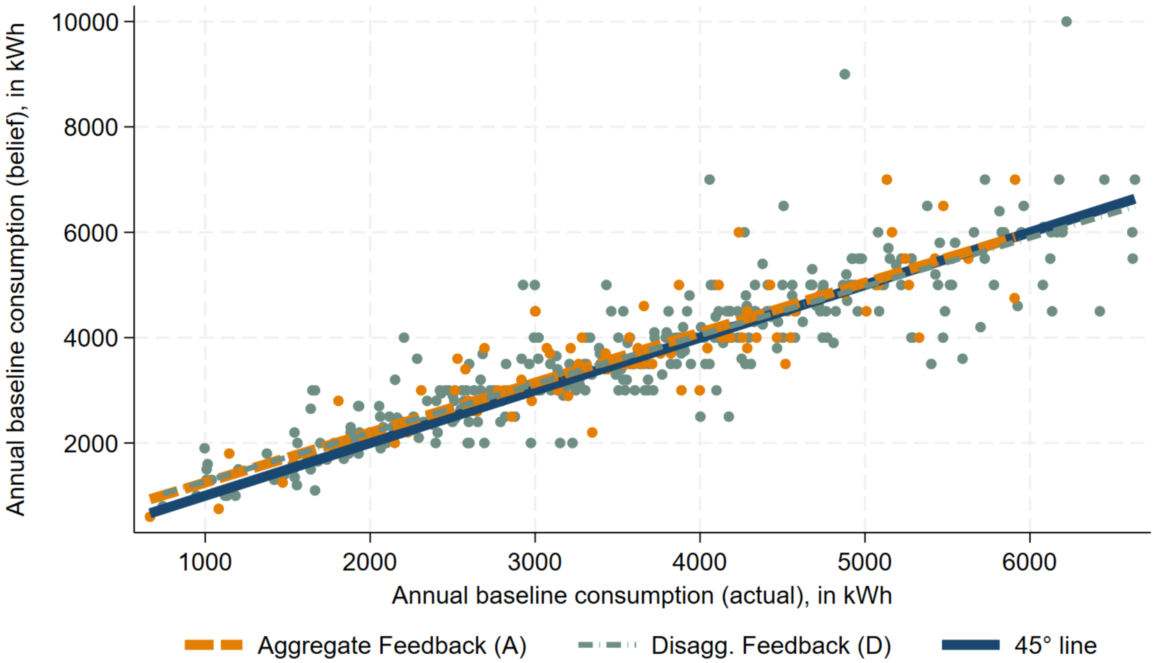
Notes: The ATE corresponds to the treatment effect of the group D (T_1-T_4), relative to group A. The outcome variable is hourly electricity consumption at the appliance level, divided by the hourly mean in the Aggregate Feedback (A) group. Shaded bars indicate that treatment effects are statistically significant at the 1% (blue shaded) or 5% (light blue shaded) level. Whiskers indicate a range of +/- 1 standard error (clustered at the household level). Based on conducting an F-test, we can reject the null hypothesis that all hourly point estimates are zero: $F(24, 699) = 1.82$, p-value $= 0.0096$.

Figure A12: ATE of Disaggregation (D) on Hourly Electricity Consumption, by Appliance Category (Relative to Aggregate Feedback (A) group)



Notes: Shaded bars indicate that treatment effects are statistically significant at the 1% (blue shaded) or 5% (light blue shaded) level. Whiskers indicate a range of +/- 1 standard error (clustered at the household level). The outcome variable is hourly electricity consumption at the appliance level, divided by the hourly mean in the Aggregate Feedback (A) group. The categories *Refrigeration* and *Always-On* are measured daily, so that we cannot estimate hourly treatment effects for them.

Figure A13: ATE of Aggregate Feedback (A) group on Hourly Electricity Consumption (relative to Matched Control)



Notes: Dots represent data points, lines stem from linear regressions of beliefs on actual baseline consumption and a constant (separately for households in the Aggregate Feedback (A) group and households in the treatment groups $T_1 - T_4$, D. We elicited beliefs about the baseline electricity consumption in a survey that we conducted prior to the field experiment. Actual baseline consumption corresponds to the consumption of a household from the last (annual) bill prior to the field experiment. We drop outliers, defined as all observations above the 95 percentile of the actual baseline consumption distribution, as well as below the 2.5 or above the 97.5 percentile of the distribution of consumption beliefs, divided by the respective actual consumption.

Supplementary Tables

Table A1: Descriptives (Experimental Sample)

	A	T_1	T_2	T_3	T_4	P-value
a) Socio-demographics (n=700)						
Baseline consumption, in kWh/day	10.2	10.5	10.1	10.5	10.5	0.90
No. of occupants	2.4	2.5	2.4	2.5	2.6	0.17
Monthly net income, in EUR	3,004	3,188	3,030	3,194	3,091	0.63
Own property, in %	73.6	79.7	73.8	77.5	73.2	0.67
Employed, in %	50.2	53.1	50.9	55.4	46.0	0.23
Share of females, in %	44.8	49.0	48.7	46.9	47.5	0.40
Age, in years	47.6	44.8	47.6	45.9	43.4	0.29
b) Baseline Aggregate Consumption Beliefs (n=466)						
Yearly consumption (belief), in kWh	3,643	3,650	3,616	3,496	3,727	0.81
Yearly consumption (actual), in kWh	3,524	3,551	3,497	3,467	3,589	0.97
c) Wattage Ranking Task (n=598)						
At least one mistake (0=no, 1=yes)	0.50	0.62	0.60	0.58	0.56	0.48
Share of mistakes in all comparisons	0.11	0.12	0.11	0.11	0.12	0.90
Number of households	140	136	143	143	138	$\sum$=700

Notes: P-values are from F-tests of mean equality in all experimental conditions (clustered at the household level). Variables are measured at the household level, except for employed, share of females, and age, which we measure at the household member level. The wattage ranking task refers to a task in the baseline survey that asked respondents to assess the wattage of a typical laptop, dish-washer, tumble dryer, and fan-heater, relative to a 100 W lightbulb on a five point Likert scale ("much lower", "lower", "about the same", "higher", "much higher"). "Share of mistakes in all comparisons" gives the total number of mistakes, divided by the number of all binary comparisons implied by that task. Aggregate consumption beliefs are elicited in the same survey. We drop observations beyond the 95 percentile of the belief distribution as well as those with extreme relative errors (below the 2.5 and above the 97.5 percentile).

Table A2: Comparison of Experimental Sample with German Population

	Experimental sample	German population
Baseline cons., in kWh/day	10.4	8.6
# of occupants	2.6	2.0
# of refrigeration appliances	2.3	2.4
Net income, in EUR per month	3,103	3,314
Own property, in %	75.3	44.0
Employed, in %	51.2	57.8
Share of females, in %	47.4	50.7
Age, in years	45.8	44.3
Years of schooling	11.0	10.5
Baseline cons., in kWh/day (1 person household)	6.0	5.5
Baseline cons., in kWh/day (2 person household)	9.9	8.8
Baseline cons., in kWh/day (3+ person household)	11.5	13.3
1 person household, in %	11.3	41.8
2 person household, in %	48.3	33.5
3+ person household, in %	40.4	24.7

Notes: German averages are taken from the following German Statistical Office publications (for the year 2017): Populalation Statistics (Mikrozensus); Environmental-Economic Accounting; Income, Receipts, and Expenditures; Consumption Expenditures. The average electricity consumption is for the baseline year 2016 (`https://www.destatis.de/DE/Themen/Gesellschaft-Umwelt/Umwelt/Materialfluesse-Energiefluesse/Tabellen/stromverbrauch-haushalte.html`, last access: March 9, 2020).

Table A3: Descriptives: Appliance Possessions and House Characteristics

	All	A	T1	T2	T3	T4	P-value	N
Appliance Possession								
# of cooling appliances	2.27	2.23	2.24	2.27	2.27	2.33	0.82	700
# of dish washers	0.92	0.91	0.91	0.95	0.93	0.91	0.87	700
# of washing machines	1.02	1.00	1.04	1.00	1.01	1.04	0.61	700
# of tumble dryers	0.72	0.74	0.71	0.69	0.69	0.77	0.68	700
# of ovens	0.98	0.98	0.99	0.98	0.99	0.98	0.76	700
# of hobs	0.99	0.99	0.99	0.98	0.99	0.99	0.84	700
HH lives in detached house (1: yes, 0: no)	0.45	0.46	0.42	0.48	0.43	0.44	0.83	700
HH lives in semi-detached house (1: yes, 0: no)	0.14	0.13	0.16	0.10	0.16	0.16	0.37	700
HH lives in an apartment (1: yes, 0: no)	0.24	0.22	0.26	0.23	0.24	0.23	0.94	700
# of bedrooms	3.03	3.06	3.01	3.00	3.00	3.08	0.88	686
Water heating via gas (1: yes, 0: no)	0.62	0.59	0.63	0.64	0.60	0.64	0.83	700
Space heating via gas (1: yes, 0: no)	0.58	0.53	0.68	0.57	0.55	0.57	0.07	700

Notes: P-values are from F-tests of mean equality in all experimental conditions: A, T_1, T_2, T_3, T_4 (heteroscedasticity-robust standard errors).

Table A4: Benchmarks for the Calculation of Efficiency Scores

	1 person	2 persons	3 persons	4 persons	5 persons
Always on	0.0	0.0	0.0	0.0	0.0
	90.7	125.7	158.7	174.1	190.5
Refrigeration (1 appliance)	4.8	4.8	5.7	5.9	6.3
	39.9	44.3	49.3	54.2	59.6
Refrigeration (2 appliances)	8.3	11.5	11.5	12.4	12.5
	66.5	73.9	82.1	90.3	99.3
Refrigeration (3+ appliances)	8.3	11.5	11.5	12.4	12.5
	94.1	103.1	112.1	121.6	132.0
Washing machine	1.2	1.8	2.3	2.8	3.4
	14.0	23.9	36.4	43.6	44.6
Dishwasher	0.0	0.0	0.0	0.0	0.0
	17.7	27.7	34.0	43.6	54.6
Dryer	0.0	0.0	0.0	0.0	0.0
	31.9	62.7	68.8	70.5	72.2
Oven	0.0	0.0	0.0	0.0	0.0
	6.3	12.6	19.0	25.3	31.6

Notes: The benchmarks were calculated taking the technical energy efficiencies of appliances on the market into account (energy efficient – energy inefficient), as well as typical usage behaviours (rare user – heavy user). The main sources for technical efficiency are product data sheets for efficient appliances from EcoTopTen, an online platform for energy efficient products (URL: `https://www.ecotopten.de/`), as well as product data sheets of inefficient appliances from product tests by Stiftung Warentest, a renowned German consumer organisation (URL: `www.test.de`). In addition, we use information on typical usage behaviours from surveys such as the German Residential Energy Consumption Survey (RWI-GRECS, URL: `http://www.rwi-essen.de/forschung-und-beratung/fdz-ruhr/datenangebot/mikrodaten/rwi-grecs`). For the category *Always-On*, we calculate the upper benchmark based on the stand-by electricity use for a range of appliances, including TVs, hifi systems, PCs, routers, telefones, coffee machines, washing machines, and microwaves (using energy inefficient appliance varieties). These calculations take typical appliance possessions by household size into account. We set the lower benchmark to zero, assuming that always-on consumption can be avoided. For the category *Refrigeration*, we calculate benchmarks based on the number of appliances (1, 2, and more than 2). For each of them, we consider the most energy efficient and inefficient appliances available on the market, whose cooling volume is as recommended for the respective household size. For the categories *Washing machine*, we use data on the energy consumption per use for energy efficient and inefficient appliances and consider the typical frequency of use for heavy users and rare users (for every household size). We proceed in the same manner for the categories *Dish washer*, *Dryer*, and *Oven*, but assume that the lower benchmark is zero as households can substitute these energy services with hand-washing, dry-hanging and eating-out, for example. Details on the calculations can be obtained from the authors upon request.

Table A5: Distribution of Daily Use, in kWh/day

	Mean	Std. dev.	Min	Max	p1	p25	p50	p75	p99	N
Total electricity	10.16	5.36	0.00	100.07	1.72	6.56	9.27	12.75	27.03	106,283
Always on	2.47	1.97	0.00	47.49	0.14	1.10	2.01	3.30	9.33	93,187
Refrigeration	1.02	0.66	0.00	11.12	0.07	0.63	0.88	1.24	3.29	93,185
Dishwasher	0.30	0.63	0.00	14.15	0.00	0.00	0.00	0.00	2.64	84,511
Washing mach.	0.49	1.11	0.00	40.30	0.00	0.00	0.00	0.60	5.08	91,473
Dryer	0.14	0.59	0.00	36.06	0.00	0.00	0.00	0.00	2.88	65,852
Oven	0.24	0.81	0.00	20.57	0.00	0.00	0.00	0.00	3.86	93,187
Other appliances	5.56	3.89	0.00	56.58	0.00	3.01	4.74	7.15	19.07	93,187

Notes: $p1$ denotes the first percentile, $p25$ the 25th percentile, etc. N denotes the number of daily appliance-level observations.

Table A6: Distribution of Use per Utilization, in kWh

	Mean	Std. dev.	Min	Max	p1	p25	p50	p75	p99	N
Always-On	2.47	1.97	0.000	47.49	0.13	1.10	2.01	3.30	9.32	93,664
Refrigeration	1.02	0.66	0.000	11.12	0.07	0.63	0.88	1.24	3.29	93,659
Dishwasher	1.08	0.42	0.002	10.98	0.34	0.79	1.04	1.32	2.37	23,643
Washing mach.	1.00	0.75	0.001	18.54	0.05	0.53	0.81	1.27	3.65	45,270
Dryer	1.02	1.02	0.004	38.46	0.03	0.38	0.74	1.37	4.51	8,979
Oven	0.80	1.11	0.001	20.09	0.01	0.19	0.47	0.95	5.55	28,362

Notes: In the categories *Always-On* and *Refrigeration*, the unit of utilization is one day. For all other appliances, average use is given per utilization event. $p1$ denotes the first percentile, $p25$ the 25th percentile, etc. N denotes the number of appliance-use events.

Table A7: Beliefs vs. Actual Baseline Consumption (at the Household-Level)

(a) Mean Annual Baseline Consumption, Beliefs vs. Actual

	Belief (in kWh)	Actual (in kWh)	Mean Diff. (in kWh)	Std. Err.	P-value	n
Pooled	3,627.6	3,539.4	-88.29	(27.74)	0.002	497
A	3,611.6	3,491.5	-120.06	(60.51)	0.048	92
T_1	3,675.8	3,614.0	-61.80	(63.01)	0.327	100
T_2	3,642.4	3,561.0	-81.34	(75.86)	0.284	99
T_3	3,531.8	3,483.0	-48.77	(44.84)	0.277	106
T_4	3,681.3	3,547.1	-134.30	(64.29)	0.037	100

(b) Alignment Between Beliefs and Actual Baseline Consumption

	(1)		(2)	
	Estimate	Std. Err.	Estimate	Std. Err.
Y^b: Baseline elec. use	0.928***	(0.026)	0.948***	(0.058)
D: T_1-T_4			48.891	(205.672)
$D \times Y^b$			-0.024	(0.065)
Constant	342.923***	(82.470)	300.114	(184.225)
R^2	0.7878		0.7879	
Number of obs.	497		497	

Notes for both Panels: Beliefs about the electricity consumption in the baseline period were elicited in a survey that we conducted prior to the field experiment. Actual baseline consumptions were obtained from billing data. ***, **,* denote statistical significance at the 1%, 5%, 10% level, respectively. Robust standard errors are in parentheses. A equals 1 for households in the Aggregate Feedback (A) group. D equals one for households in the groups $T_1 - T_4$. We drop outliers, defined as all observations above the 95 percentile of the actual baseline consumption, as well as below the 2.5 or above the 97.5 percentile of the consumption belief, divided by the actual consumption.
Notes for Panel b): In Column 1), we pool all experimental groups and estimate the model: $Y^{b,belief} = \alpha + \beta Y^b + \epsilon$, where $Y^{b,belief}$ and Y^b denotes the belief and the actual consumption of household i in the baseline period, respectively (and ϵ denotes an error term). In Column (2), we interact βY^b with the disaggregation dummy D that equals one for the groups T$_1$-T$_4$: $Y^{b,belief} = \alpha + \beta Y^b + \gamma D \cdot Y^b + \epsilon$. The estimate $\hat{\beta}$ gives the average change in beliefs as the actual consumption increases by one unit (for the *EC* group). The estimate $\hat{\gamma}$ gives the the change of this slope for the households in the treatment groups T$_1$-T$_4$ (relative to the *EC* group).

Table A8: Balance Experimental Sample (Group A) vs. Matched Non-Experimental Sample

	Difference A-MC	Std. Err.	P-val.	N
Baseline elec. cons., in kWh/day	0.23	0.27	0.40	280
End of baseline billing, in days	−10.04	11.58	0.39	280
Start of baseline billing, in days	−33.75	13.28	0.01	280

Notes: ***, **,* denote statistical significance at the 1%, 5%, 10% level, respectively. Standard errors in parentheses, clustered at the household level.

Table A9: Heterogeneity in ATE, by Baseline Consumption

(a) Experimental Sample

	Estimate	Std. Err.
D: Disaggregation	-0.051^{***}	(0.017)
$D \times Y^{b,dm}$	-0.066	(0.052)
$Y^{b,dm}$: Baseline elec. use (demeaned)	0.939^{***}	(0.045)
Day fixed effects	✓	
Month-of-baseline FE	✓	
R^2	0.5480	
Number of obs.	106,283	
Number of households	700	

(b) Sample Including Matched Control

	Estimate	Std. Err.
D: Disaggregation	-0.047^{***}	(0.015)
$D \times Y^{b,dm}$	-0.134^{***}	(0.044)
$Y^{b,dm}$	1.007^{***}	(0.035)
A: Exp. Control	0.008	(0.021)
$A \times Y^{b,dm}$	-0.068	(0.058)
Day fixed effects	✓	
Month-of-baseline FE	✓	
R^2	0.5634	
Number of obs.	127,790	
Number of households	840	

Notes: ***, **,* denote statistical significance at the 1%, 5%, 10% level, respectively. Standard errors are in parentheses and clustered at the household and the household-match level for Panel a) and b), respectively. The outcome variable is daily electricity consumption, divided by the mean in the Aggregate Feedback (A) group. Baseline electricity use $Y^{b,dm}$ is demeaned.

Table A10: ATEs on Daily Electricity Consumption, Relative to Aggregate Feedback (A) group (Balanced Panel)

(a) Effect of Experimental Conditions

	(1)	(2)
D: $T_1 - T_4$	-0.064***	
	(0.019)	
T_1		-0.065***
		(0.023)
T_2		-0.048**
		(0.022)
T_3		-0.087***
		(0.024)
T_4		-0.056**
		(0.027)
Y^b: Baseline elec. use	0.878***	0.879***
	(0.028)	(0.028)
Day fixed effects (FE)	✓	✓
Month-of-baseline FE	✓	✓
R^2	0.5373	0.5380
Number of obs.	79,562	79,562
Number of households	460	460

(b) Effects of App Elements

	(3)	(4)	(5)
D: Disaggregation	-0.064***	-0.068***	-0.064***
	(0.019)	(0.021)	(0.022)
M: Monetary incentives		0.025	0.015
		(0.015)	(0.017)
R: Ranking		-0.016	-0.026
		(0.015)	(0.020)
M: Monet. inc. $\times$ R: Rank.			0.021
			(0.030)
Y^b: Baseline elec. use	0.878***	0.880***	0.879***
	(0.028)	(0.028)	(0.028)
Day fixed effects (FE)	✓	✓	✓
Month-of-baseline FE	✓	✓	✓
R^2	0.5373	0.5379	0.5380
Number of obs.	79,562	79,562	79,562
Number of households	460	460	460

Notes: ***, **,* denote statistical significance at the 1%, 5%, 10% level, respectively. The regressions are based on a balanced panel for the core study period (months 1-6). Standard errors are in parantheses and clustered at the household-match level. The outcome variable is daily electricity consumption, divided by the mean in the Aggregate Feedback (A) group. D equals one for households in the groups $T_1 - T_4$, M equals one for households in the groups T_2 and T_4, and R equals one for households in the groups T_3 and T_4, while being zero for the other participants, respectively.

Table A11: ATEs on Daily Electricity Consumption, Relative to Matched Control (MC) group

(a) Effect of Experimental Conditions

	(1)	(2)
A:	-0.009	-0.009
	(0.022)	(0.022)
D: $T_1 - T_4$	-0.055***	
	(0.016)	
T_1		-0.053***
		(0.019)
T_2		-0.049**
		(0.020)
T_3		-0.073***
		(0.021)
T_4		-0.047**
		(0.021)
Y^b: Baseline elec. use	0.910***	0.910***
	(0.020)	(0.020)
Day fixed effects	✓	✓
Month-of-baseline FE	✓	✓
R^2	0.5687	0.5689
Number of obs.	127,790	127,790
Number of households	840	840

(b) Effects of App Elements

	(3)	(4)	(5)
AF: Aggregate feedback	-0.009	-0.009	-0.009
	(0.022)	(0.022)	(0.022)
D: Disaggregation	-0.047***	-0.049***	-0.045**
	(0.016)	(0.018)	(0.018)
M: Monetary incentives		0.016	0.005
		(0.013)	(0.016)
R: Ranking		-0.010	-0.021
		(0.013)	(0.018)
M: Monet. inc. $\times$ R: Rank.			0.023
			(0.027)
Y^b: Baseline elec. use	0.910***	0.910***	0.910***
	(0.020)	(0.020)	(0.020)
Day fixed effects	✓	✓	✓
Month-of-baseline FE	✓	✓	✓
R^2	0.5687	0.5689	0.5690
Number of obs.	127,790	127,790	127,790
Number of households	840	840	840

Notes: ***, **,* denote statistical significance at the 1%, 5%, 10% level, respectively. These regressions include observations from our Matched Control (MC) group, as described in Section A4. Standard errors are in parentheses and clustered at the household-match level. The outcome variable is daily electricity consumption, divided by the mean in the Aggregate Feedback (A) group. A equals one for the households in the groups A and $T_1 - T_4$, while being zero for households in MC. D equals one for households in the groups $T_1 - T_4$, M equals one for households in the groups T_2 and T_4, and R equals one for households in the groups T_3 and T_4, while being zero for the other participants, respectively.

Table A12: ATE on Daily Electricity Consumption (After the Core Study Period), Only Main Effects

(a) Experimental Sample

	(1) Full Sample	(2) Balanced Panel
D: Disaggregation	-0.014	-0.033
	(0.017)	(0.022)
Y^b: Baseline elec. use	0.747***	0.732***
	(0.025)	(0.034)
Day fixed effects	✓	✓
Month-of-baseline FE	✓	✓
R^2	0.4978	0.4909
Number of obs.	54,603	29,330
Number of households	586	321

(b) Experimental and Matched Control Samples

	(3) Full Sample	(4) Balanced Panel
A: Aggregate Feedback	-0.036	-0.026
	(0.026)	(0.035)
D: Disaggregation	-0.047**	-0.027
	(0.022)	(0.023)
Y^b: Baseline elec. use	0.781***	0.760***
	(0.024)	(0.032)
Day fixed effects	✓	✓
Month-of-baseline FE	✓	✓
R^2	0.5112	0.5004
Number of obs.	66,084	35,765
Number of households	708	391

Notes: ***, **,* denote statistical significance at the 1%, 5%, 10% level, respectively. Standard errors are in parentheses and clustered at the household level in Columns (1) and (2), and at the household-match level for Columns (3) and (4). The outcome variable is daily electricity consumption, divided by the mean in the Aggregate Feedback (A) group. D equals one for households in the groups $T_1 - T_4$, while being zero for other participants.

Table A13: Difference in ATEs During and After the Core Study Period

	(1) Experimental Sample	(2) Experimental and Matched Control Samples
1(After Core Study Period)	-0.052*	-0.008
	(0.031)	(0.034)
A : Aggregated Feeback (A)		-0.010
		(0.032)
A : Aggregated Feeback $\times$ 1(After Core Study Period)		-0.025
		(0.033)
D: Disaggregation	-0.063**	-0.060**
	(0.025)	(0.026)
D: Disaggregation $\times$ 1(After Core Study Period)	0.038	0.038
	(0.023)	(0.023)
Y^b: Baseline elec. use	0.840***	0.866***
	(0.032)	(0.028)
R^2	0.5280	0.5450
Number of obs.	84,891	103,411
Number of participants	321	391

Notes: ***, **,* denote statistical significance at the 1%, 5%, 10% level, respectively. Standard errors are in parentheses and clustered at the household and at the household-match level for Columns (1) and (2), respectively. The outcome variable is daily electricity consumption, divided by the mean in the Aggregate Feedback (A) group. 1(After Core Study Period) is a dummy variable that equals one when an observation occurs from month 7 onwards, while being zero otherwise. D equals one for households in the groups $T_1 - T_4$, while being zero for other participants.

Table A14: Difference in ATEs, by Subgroups

	(1)	(2)
D: Disaggregation	-0.040	-0.019
	(0.028)	(0.028)
1(Two Occupant)	0.071**	
	(0.030)	
1(Three+ Occupants)	0.128***	
	(0.036)	
1(Own Property)		0.056*
		(0.029)
D: Disaggregation $\times$ 1(Two Occupant)	0.003	
	(0.035)	
D: Disaggregation $\times$ 1(Three+ Occupants)	-0.035	
	(0.039)	
D: Disaggregation $\times$ 1(Own Property)		-0.040
		(0.034)
Y^b: Baseline elec. use	0.864***	0.888***
	(0.024)	(0.024)
R^2	0.5614	0.5598
Number of obs.	106,283	105,846
Number of participants	700	696

Notes: ***, **,* denote statistical significance at the 1%, 5%, 10% level, respectively. Standard errors are in parentheses and clustered at the household level. The outcome variable is daily electricity consumption, divided by the mean in the Aggregate Feedback (A) group. 1(Two Occupants) is a dummy variable that equals one when a household consists of two occupants, while 1(Three+ Occupants) is one for households with at least three occupants. 1(Own Property) equals one if a household lives in his own property. D equals one for households in the groups $T_1 - T_4$, while being zero for other participants.

Table A15: Sensitivity of Welfare Effects

ω	η	dCS	$dCS_{always-on}$	$dCS_{cooling}$	$dCS_{dish-washer}$	$dCS_{washing}$	dCS_{dryer}	dCS_{oven}	$dCS_{residual}$	dCS_{agg}^{min}	dCS_{agg}^{max}
0.0	-0.4400	6.9405	0.0005	0.0056	0.2615	0.0380	3.6002	0.0138	3.0209	0.3092	7.4976
0.1	-0.4354	5.5930	0.0006	0.0056	0.2638	0.0371	2.1456	0.0121	3.1282	0.3186	4.4683
0.2	-0.4308	5.0911	0.0006	0.0057	0.2662	0.0362	1.5281	0.0108	3.2435	0.3286	3.1825
0.3	-0.4261	4.8742	0.0006	0.0057	0.2686	0.0353	1.1867	0.0097	3.3676	0.3393	3.0761
0.4	-0.4215	4.7923	0.0006	0.0058	0.2711	0.0345	0.9699	0.0088	3.5016	0.3506	3.1043
0.5	-0.4169	4.7886	0.0006	0.0058	0.2736	0.0337	0.8201	0.0081	3.6466	0.3628	3.1331
0.6	-0.4123	4.8378	0.0007	0.0059	0.2762	0.0330	0.7104	0.0075	3.8042	0.3758	3.1625
0.7	-0.4077	4.9272	0.0007	0.0059	0.2788	0.0323	0.6266	0.0069	3.9760	0.3898	3.1924
0.8	-0.4030	5.0508	0.0007	0.0060	0.2814	0.0316	0.5605	0.0065	4.1641	0.4049	3.2228
0.9	-0.3984	5.2058	0.0007	0.0060	0.2842	0.0309	0.5070	0.0061	4.3709	0.4212	3.2539
1.0	-0.3938	5.3919	0.0008	0.0061	0.2869	0.0303	0.4628	0.0057	4.5993	0.4389	3.2855
1.1	-0.3892	5.6103	0.0008	0.0062	0.2897	0.0297	0.4257	0.0054	4.8528	0.4581	3.3178
1.2	-0.3846	5.8640	0.0008	0.0062	0.2926	0.0291	0.3941	0.0051	5.1360	0.4790	3.3507
1.3	-0.3800	6.1572	0.0009	0.0063	0.2955	0.0286	0.3669	0.0049	5.4542	0.5020	3.3842
1.4	-0.3753	6.4961	0.0009	0.0064	0.2985	0.0280	0.3432	0.0046	5.8145	0.5273	3.4185
1.5	-0.3707	6.8890	0.0010	0.0064	0.3016	0.0275	0.3223	0.0044	6.2257	0.5553	3.4534

Notes: We test the sensitivity of our welfare analysis for different appliance-level elasticities, holding the aggregate elasticity roughly constant between 0.39 (as implied by our estimates) and $\eta^F = 0.44$ (as estimated by Frondel et al. 2019). In particular, we investigate the following appliance-level elasticities $\hat{\eta}_j = \omega\eta_j + (1-\omega)\eta^F$. This specification ensures that a) aggregate elasticities are in line with previous evidence (Frondel et al., 2019) , and b) that the heterogeneity in appliance-level biases increases as ω increases. *dCS* denotes the change in consumer surplus from disaggregate feedback (subscripts denote appliance-specific effects. dCS_{agg}^{min} and dCS_{agg}^{max} denote our bounds for the consumer surplus effects of aggregate feedback.

Table A16: Effect Heterogeneity and Appliance-Level Beliefs

(a) Effect Heterogeneity by Wattage Ranking Task Results

	(1) Estimate	(1) Std. err.	(2) Estimate	(2) Std. err.	(3) Estimate	(3) Std. err.
1(At least one mistake)	0.028	(0.021)	0.031	(0.022)		
1(No wattage ranking)			0.169**	(0.068)		
$D \times$ 1(No mistake)	−0.014	(0.019)	−0.010	(0.020)		
$D \times$ 1(At least one mistake)	−0.043**	(0.018)	−0.043**	(0.018)		
$D \times$ 1(No wattage ranking)			−0.155**	(0.069)		
D: Disaggregation					−0.012	(0.017)
Share of mistakes					0.125*	(0.067)
$D \times$ Share of mistakes					−0.156*	(0.085)
Y^b: Baseline elec. use	0.888***	(0.024)	0.893***	(0.022)	0.889***	(0.024)
Day fixed effects (FE)	✓	✓	✓	✓	✓	✓
Month-of-baseline FE	✓	✓	✓	✓	✓	✓
R^2	0.5613		0.5613		0.5615	
Number of obs.	91,638		106,283		91,638	
Number of participants	598		700		598	

(b) Appliance-Level Rank Differences between Beliefs and Measured Uses

	(1) Mean rank diff. (B)	(2) Mean rank diff. (E)	(3) Change (E−B)	(4) Std. err.	(5) P-value	(6) N
A	1.81	1.79	−0.02	0.07	0.757	66
D: T_1-T_4	1.76	1.56	−0.19	0.04	0.000	296
T_1	1.75	1.57	−0.18	0.10	0.077	61
T_2	1.75	1.42	−0.32	0.08	0.000	62
T_3	1.69	1.47	−0.22	0.10	0.031	54
T_4	1.79	1.54	−0.26	0.09	0.009	53

Notes: ***, **,* denote statistical significance at the 1%, 5%, 10% level, respectively. Standard errors are in parentheses and clustered at the household level. "1(At least one mistake)" and "1(No wattage ranking)" denote dummy variables that equal one if a household made at least one mistake in the wattage ranking task or did not take part in that task, respectively. "Mean rank diff." denotes the mean of the absolute rank differences between the rank of appliance consumptions (measured during the study period) and the rank of electricity consumption beliefs, which we elicited for every appliance category before (baseline, B) and during the study period (endline, E).

Table A17: Appliance-Level Elasticity Estimates

	(1) Always-On	(2) Refrigeration	(3) Dish-Washer	(4) Washing	(5) Dryer	(6) Oven	(7) Other appl.
ln p	-0.310	-0.400*	-0.401	-0.552	-3.423***	-1.067	-0.289
	(0.355)	(0.226)	(0.722)	(0.418)	(1.233)	(1.024)	(0.252)
# of occupants = 2	0.113	0.129***	0.872***	0.843***	-0.012	0.694***	0.278***
	(0.069)	(0.043)	(0.192)	(0.129)	(0.359)	(0.200)	(0.055)
# of occupants = 3	0.038	0.116**	1.099***	0.937***	0.104	0.970***	0.301***
	(0.084)	(0.052)	(0.201)	(0.140)	(0.396)	(0.229)	(0.058)
# of occupants = 4	0.126	0.098*	1.178***	1.019***	0.135	1.104***	0.321***
	(0.087)	(0.058)	(0.227)	(0.144)	(0.406)	(0.245)	(0.062)
# of occupants = 5	-0.071	0.212***	1.460***	0.692***	0.508	0.947***	0.264***
	(0.106)	(0.074)	(0.287)	(0.191)	(0.422)	(0.325)	(0.081)
Hh. net income	0.048***	-0.017*	-0.001	0.026	-0.026	-0.042	-0.001
	(0.013)	(0.009)	(0.035)	(0.021)	(0.052)	(0.037)	(0.008)
1(Hh. net income missing)	0.208**	-0.100	-0.038	0.126	-0.419	-0.032	0.031
	(0.091)	(0.064)	(0.241)	(0.140)	(0.363)	(0.252)	(0.061)
Y^b: Baseline elec. use	0.106***	0.030***	0.049***	0.059***	0.044***	0.072***	0.082***
	(0.006)	(0.004)	(0.011)	(0.009)	(0.014)	(0.013)	(0.005)
Constant	-1.230**	-0.986***	-3.728***	-3.338***	-7.236***	-4.725***	0.058
	(0.507)	(0.326)	(1.066)	(0.616)	(1.879)	(1.504)	(0.356)
R^2	0.489	0.169	0.154	0.288	0.092	0.151	0.635
Number of participants	700	700	610	677	259	527	700

Notes: ***, **,* denote statistical significance at the 1%, 5%, 10% level, respectively. The outcome variable is the average daily electricity consumption in the study period (in logs). The regressors are net income, baseline electricity use, as well as a set of dummy variables for the number of occupants and missings for the variable net income. Heteroscedasticity robust standard errors in parentheses.

References

Abadie, Alberto, and Jann Spiess. 2021. "Robust Post-Matching Inference." *Journal of the American Statistical Association* 1–13. 10.1080/01621459.2020.1840383.

Allcott, Hunt. 2011. "Social Norms and Energy Conservation." *Journal of Public Economics* 95 (9-10): 1082–1095.

Allcott, Hunt. 2015. "Site Selection Bias in Program Evaluation." *The Quarterly Journal of Economics* 130 (3): 1117–1165.

Allcott, Hunt, and Dmitry Taubinsky. 2015. "Evaluating Behaviorally Motivated Policy: Experimental Evidence from the Lightbulb Market." *American Economic Review* 105 (8): 2501–2538.

Andor, Mark A, Andreas Gerster, Jörg Peters, and Christoph M Schmidt. 2020a. "Social Norms and Energy Conservation Beyond the US." *Journal of Environmental Economics and Management* 103 102351.

Andor, Mark A., Andreas Gerster, and Stephan Sommer. 2020b. "Consumer Inattention, Heuristic Thinking and the Role of Energy Labels." *The Energy Journal* 41 (1): 83–122.

Asensio, Omar I., and Magali A. Delmas. 2015. "Nonprice Incentives and Energy Conservation." *Proceedings of the National Academy of Sciences* 112 (6): 510–515.

Attari, Shahzeen Z., Michael L. DeKay, Cliff I. Davidson, and Wändi Bruine de Bruin. 2011. "Reply to Frederick et al.: Anchoring Effects on Energy Perceptions." *Proceedings of the National Academy of Sciences* 108 (8): E24–E24.

Attari, Shahzeen Z., Michael L. DeKay, Cliff I. Davidson, and de Bruin, Wändi Bruine. 2010. "Public Perceptions of Energy Consumption and Savings." *Proceedings of the National Academy of Sciences* 107 (37): 16054–16059.

Azmat, Ghazala, and Nagore Iriberri. 2010. "The Importance of Relative Performance Feedback Information: Evidence from a Natural Experiment using High School Students." *Journal of Public Economics* 94 (7-8): 435–452.

Becker, Gary S. 1965. "A Theory of the Allocation of Time." *Economic Journal* 75 (299): 493–517.

Benabou, Roland, and Jean Tirole. 2002. "Self-Confidence and Personal Motivation." *Quarterly Journal of Economics* 117 (3): 871–915. 10.1162/003355302760193913.

Bernheim, B Douglas, and Dmitry Taubinsky. 2018. "Behavioral Public Economics." *Handbook of Behavioral Economics: Applications and Foundations 1* 1 381–516.

Bollinger, Bryan, Phillip Leslie, and Alan Sorensen. 2011. "Calorie Posting in Chain Restaurants." *American Economic Journal: Economic Policy* 3 (1): 91–128.

Brent, Daniel A., and Michael B. Ward. 2019. "Price Perceptions in Water Demand." *Journal of Environmental Economics and Management* 98 102266.

Bruelisauer, Marcel, Lorenz Goette, Zhengyi Jiang, Jan Schmitz, and Renate Schubert. 2018. "Appliance Specific Feedback and Social Comparisons: Evidence From a Field Experiment on Electricity Saving." *SSRN Working Paper*. 10.2139/ssrn.3299500.

Burkhardt, Jesse, Kenneth Gillingham, and Praveen K Kopalle. 2019. "Experimental Evidence on the Effect of Information and Pricing on Residential Electricity Consumption." Working Paper 25576, National Bureau of Economic Research.

Byrne, David, Lorenz Goette, Leslie Martin, Samuel Schoeb, Verena Tiefenbeck, and Thorsten Staake. 2020. "The Behavioral Mechanisms of Habit Formation: Evidence from a Field Experiment."Technical report, University of Melbourne.

Cadmus-Bertram, Lisa A., Bess H. Marcus, Ruth E. Patterson, Barbara A. Parker, and Brittany L. Morey. 2015. "Randomized Trial of a Fitbit-Based Physical Activity Intervention for Women." *American Journal of Preventive Medicine* 49 (3): 414–418.

Carrie Armel, K., Abhay Gupta, Gireesh Shrimali, and Adrian Albert. 2013. "Is disaggregation the holy grail of energy efficiency? The case of electricity." *Energy Policy* 52 213–234. 10.1016/j.enpol.2012.08.062.

Carroll, James, Seán Lyons, and Eleanor Denny. 2014. "Reducing Household Electricity Demand Through Smart Metering: The Role of Improved Information about Energy Saving." *Energy Economics* 45 234–243.

Chetty, Raj. 2009. "Sufficient statistics for welfare analysis: A bridge between structural and reduced-form methods." *Annu. Rev. Econ.* 1 (1): 451–488.

Chetty, Raj, Adam Looney, and Kory Kroft. 2009. "Salience and Taxation: Theory and Evidence." *American Economic Review* 99 (4): 1145–1177.

Chorus, Caspar G., Eric J. E. Molin, and Bert van Wee. 2006. "Use and Effects of Advanced Traveller Information Services (ATIS): A Review of the Literature." *Transport Reviews* 26 (2): 127–149.

Darby, Sarah. 2006. "The Effectiveness of Feedback on Energy Consumption." *A Review for DEFRA of the Literature on Metering, Billing and direct Displays* 486 (2006): 26.

Degen, Kathrin, Charles Efferson, Fabian Frei, Lorenz Goette, and Rafael Lalive. 2013. "Smart Metering, Beratung oder Sozialer Vergleich: Was Beeinflusst den Elektrizitätsverbrauch." Final report to the Swiss Federal Office of Energy.

Donnal, J. S., and S. B. Leeb. 2015. "Noncontact Power Meter." *IEEE Sensors Journal* 15 (2): 1161–1169.

EC. 2014a. "Benchmarking Smart Metering Deployment in the EU-27 with a Focus on Electricity." `https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=COM:2014:356:FIN`, Report by the European Commission.

EC. 2014b. "Cost-Benefit Analyses & State of Play of Smart Metering Deployment in the EU-27.." https://eur-lex.europa.eu/legal-content/EN/HIS/?uri=SWD:2014:189:FIN, European Commission Staff Working Document.

Enke, Benjamin, and Thomas Graeber. 2023. "Cognitive uncertainty." *The Quarterly Journal of Economics* 138 (4): 2021–2067.

Fang, Ximeng, Lorenz Goette, Bettina Rockenbach, Matthias Sutter, Verena Tiefenbeck, Samuel Schoeb, Thorsten Staake et al. forthcoming. "Complementarities in Behavioral Interventions: Evidence from a Field Experiment on Energy Conservation." *Journal of Public Economics*.

Farhi, Emmanuel, and Xavier Gabaix. 2020. "Optimal Taxation with Behavioral Agents." *American Economic Review* 110 (1): 298–336.

Faruqui, Ahmad, Doug Mitarotonda, Lisa Wood, Adam Cooper, and Judith Schwartz. 2011. "The Costs and Benefits of Smart Meters for Residential Customers." *IAEE Whitepaper*, https://www.edisonfoundation.net/iee/Documents/IEE_BenefitsofSmartMeters_Final.pdf.

Frederick, Shane W., Andrew B. Meyer, and Daniel Mochon. 2011. "Characterizing Perceptions of Energy Consumption." *Proceedings of the National Academy of Sciences* 108 (8): E23–E23.

Frondel, Manuel, Gerhard Kussel, and Stephan Sommer. 2019. "Heterogeneity in the Price Response of Residential Electricity Demand: A Dynamic Approach for Germany." *Resource and Energy Economics* 57 119–134.

Gabaix, Xavier. 2017. "Behavioral Inattention." *National Bureau of Economic Research Working Paper* No. 24096.

Gerster, Andreas, and Michael Kramm. 2019. "Correct Me If You Can - Optimal Non-Linear Taxation of Internalities." *CESifo Working Paper* No. 7732.

Giordano, Vincenzo, Ijeoma Onyeji, Gianluca Fulli, M. Jiménez, and Constantina Filiou. 2012. "Guidelines for Cost Benefit Analysis of Smart Metering Deployment." 10.2790/39435, European Commission Joint Research Centre.

Gosnell, Greer, Ralf Martin, Mirabelle Muûls, Quentin Coutellier, Goran Strbac, Mingyang Sun, and Simon Tindermans. 2019. "Making Smart Meters Smarter the Smart Way." *CEP Discussion Paper* No. 1602.

Gupta, Abahy, Vivek Garud, Uri Rodney, and Subodh Bapat. 2017. "Systems and Methods for Improving the Accuracy of Appliance Level Disaggregation in Non-Intrusive Appliance Load Monitoring Techniques.." US Patent No: US 9,612.286 B2.

Gupta, Sidhant, Matthew S. Reynolds, and Shwetak N. Patel. 2010. "ElectriSense: Single-point Sensing Using EMI for Electrical Event Detection and Classification in the Home." In

Proceedings of the 12th ACM International Conference on Ubiquitous Computing, UbiComp '10 139–148, New York, NY, USA: ACM.

Hart, George William. 1992. "Nonintrusive Appliance Load Monitoring." *Proceedings of the IEEE* 80 (12): 1870–1891.

Houde, Sébastien, Annika Todd, Anant Sudarshan, June A. Flora, and K. Carrie Armel. 2013. "Real-Time Feedback and Electricity Consumption: A Field Experiment Assessing the Potential for Savings and Persistence." *The Energy Journal* 34 (1): 87–102.

Hsiang, Solomon. 2016. "Climate econometrics." *Annual Review of Resource Economics* 8 43–75.

IEA. 2017. "Smart Meter Deployment in Key Markets, 2013-2017." `https://www.iea.org/data-and-statistics/charts/smart-meter-deployment-in-key-markets-2013-2017`, International Energy Agency.

Imbens, Guido W., and Jeffrey M. Wooldridge. 2009. "Recent Developments in the Econometrics of Program Evaluation." *Journal of Economic Literature* 47 (1): 5–86.

Ito, Koichiro, Takanori Ida, and Makoto Tanaka. 2018. "Moral Suasion and Economic Incentives: Field Experimental Evidence from Energy Demand." *American Economic Journal: Economic Policy* 10 (1): 240–267.

Jensen, Robert. 2010. "The (Perceived) Returns to Education and the Demand for Schooling." *Quarterly Journal of Economics* 125 (2): 515–548.

Karlin, Beth, Joanne F. Zinger, and Rebecca Ford. 2015. "The Effects of Feedback on Energy Conservation: A Meta-Analysis." *Psychological Bulletin* 141 (6): 1205–1227.

Martin, Steve, and Nicholas Rivers. 2017. "Information Provision, Market Incentives, and Household Electricity Consumption: Evidence from a Large-Scale Field Deployment." *Journal of the Association of Environmental and Resource Economists* 5 (1): 207–231.

Mas, Alexandre, and Enrico Moretti. 2009. "Peers at Work." *American Economic Review* 99 (1): 112–45.

McKerracher, Colin, and Jacopo Torriti. 2013. "Energy Consumption Feedback in Perspective: Integrating Australian Data to Meta-Analyses on In-Home Displays." *Energy Efficiency* 6 (2): 387–405.

O'Donoghue, Ted, and Matthew Rabin. 2006. "Optimal Sin Taxes." *Journal of Public Economics* 90 (10-11): 1825–1849.

Pop-Eleches, Cristian, Harsha Thirumurthy, James P. Habyarimana, Joshua G. Zivin, Markus P. Goldstein, Damien de Walque, Leslie MacKeen, Jessica Haberer, Sylvester Kimaiyo, John Sidle, Duncan Ngare, and David R. Bangsberg. 2011. "Mobile Phone Technologies Improve Adherence to Antiretroviral Treatment in a Resource-Limited Setting: a Randomized Controlled Trial of Text Message Reminders." *AIDS* 25 (6): 825–834.

Rees-Jones, Alex, and Dmitry Taubinsky. 2019. "Measuring "Schmeduling"." *The Review of Economic Studies* 87 (5): 2399–2438.

Schleich, Joachim, Corinne Faure, and Marian Klobasa. 2017. "Persistence of the Effects of Providing Feedback Alongside Smart Metering Devices on Household Electricity Demand." *Energy Policy* 107 225–233.

Tiefenbeck, Verena, Lorenz Goette, Kathrin Degen, Vojkan Tasic, Elgar Fleisch, Rafael Lalive, and Thorsten Staake. 2018. "Overcoming Salience Bias: How Real-Time Feedback Fosters Resource Conservation." *Management Science* 64 (3): 983–1476.

Tiefenbeck, Verena, Anselma Wörner, Samuel Schöb, Elgar Fleisch, and Thorsten Staake. 2019. "Real-Time Feedback Promotes Energy Conservation in the Absence of Volunteer Selection Bias and Monetary Incentives." *Nature Energy* 4 (1): 35–41.

Tran, Anh, and Richard Zeckhauser. 2012. "Rank as an Inherent Incentive: Evidence from a Field Experiment." *Journal of Public Economics* 96 (9-10): 645–650.

Blanes i Vidal, Jordi, and Mareike Nossol. 2011. "Tournaments without Prizes: Evidence from Personnel Records." *Management Science* 57 (10): 1721–1736.

Wichman, Casey J. 2017. "Information Provision and Consumer Behavior: A Natural Experiment in Billing Frequency." *Journal of Public Economics* 152 13–33.